

Statistical and computational trade-offs in estimation of sparse principal components

Tengyao Wang, Quentin Berthet, and Richard J. Samworth

University of Cambridge, California Institute of Technology
and University of Cambridge

(August 25, 2014)

Abstract

In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. In this paper we show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a Restricted Covariance Concentration condition, we show that no randomised polynomial time algorithm can achieve the minimax optimal rate. On the other hand, we also study a (polynomial time) variant of the well-known semidefinite relaxation estimator, and show that it attains essentially the optimal rate among all randomised polynomial time algorithms.

1 Introduction

Principal Component Analysis (PCA), which involves projecting a sample of multivariate data onto the space spanned by the leading eigenvectors of the sample covariance matrix,

is one of the oldest and most widely-used dimension reduction devices in Statistics. It has proved to be particularly effective when the dimension of the data is relatively small by comparison with the sample size. However, the work of [Johnstone and Lu \(2009\)](#) and [Paul \(2007\)](#) shows that PCA breaks down in the high-dimensional settings that are frequently encountered in many diverse modern application areas. For instance, consider the spiked covariance model where X_1, \dots, X_n are independent $N_p(0, \Sigma)$ random vectors, with $\Sigma = I_p + \theta v_1 v_1^\top$ for some $\theta > 0$ and an arbitrary unit vector $v_1 \in \mathbb{R}^p$. In this case, v_1 is the leading eigenvector (principal component) of Σ , and the classical PCA estimate would be \hat{v}_1 , a unit-length leading eigenvector of the sample covariance matrix $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$. In the high-dimensional setting where $p = p_n$ is such that $p/n \rightarrow c \in (0, 1)$, [Paul \(2007\)](#) showed that

$$|\hat{v}_1^\top v_1| \xrightarrow{a.s.} \begin{cases} \sqrt{\frac{1-c/\theta^2}{1+c/\theta}} & \text{if } \theta > \sqrt{c} \\ 0 & \text{if } \theta \leq \sqrt{c}. \end{cases}$$

In other words, \hat{v}_1 is inconsistent as an estimator of v_1 in this asymptotic regime. This phenomenon is related to the so-called ‘BBP’ transition in random matrix theory ([Baik, Ben Arous and P     , 2005](#)).

Sparse Principal Component Analysis was designed to remedy this inconsistency and to give additional interpretability to the projected data. In the simplest case, it is assumed that the leading eigenvector v_1 of the population covariance matrix Σ belongs to the k -sparse unit Euclidean sphere in \mathbb{R}^p , given by $B_0(k) := \{u = (u_1, \dots, u_p)^\top \in \mathbb{R}^p : \sum_{j=1}^p \mathbb{1}_{\{u_j \neq 0\}} \leq k, \|u\|_2 = 1\}$. A remarkable number of recent papers have proposed estimators of v_1 in this setting, including [Jolliffe, Trendafilov and Uddin \(2003\)](#), [Zou, Hastie and Tibshirani \(2006\)](#), [d’Aspremont et al. \(2007\)](#), [Johnstone and Lu \(2009\)](#), [Witten, Tibshirani and Hastie \(2009\)](#), [Journ  e et al. \(2010\)](#), [Birnbaum et al. \(2013\)](#), [Cai, Ma and Wu \(2013\)](#), [Ma \(2013\)](#), [Shen, Shen and Marron \(2013\)](#) and [Vu and Lei \(2013\)](#). In [Birnbaum et al. \(2013\)](#), [Ma \(2013\)](#) and [Shen, Shen and Marron \(2013\)](#), the authors were able to show that their estimators attain the minimax rate of convergence over certain Gaussian classes of distributions, provided that k is treated as a fixed constant. Both [Cai, Ma and Wu \(2013\)](#) and [Vu and Lei \(2013\)](#) also study minimax properties, but treat k as a parameter of the problem that may vary with the sample size n . In particular, for a certain class $\mathcal{P}_p(n, k)$ of subgaussian distributions and

in a particular asymptotic regime, [Vu and Lei \(2013\)](#) show¹ that

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k)} \mathbb{E}_P \{1 - (v_1^\top \hat{v})^2\} \asymp \frac{k \log p}{n},$$

where the infimum is taken over all estimators \hat{v} . Moreover, they show that the minimax rate is attained by a leading k -sparse eigenvector of $\hat{\Sigma}$, given by

$$\hat{v}_{\max}^k \in \operatorname{argmax}_{u \in B_0(k)} u^\top \hat{\Sigma} u. \quad (1)$$

The papers cited above would appear to settle the question of sparse principal component estimation (at least in a subgaussian setting) from the perspective of statistical theory. However, there remains an unsettling feature, namely that neither the estimator of [Cai, Ma and Wu \(2013\)](#), nor that of [Vu and Lei \(2013\)](#), is computable in polynomial time². For instance, computing the estimator (1) is an NP-hard problem, and the naive algorithm that searches through all $\binom{p}{k}$ of the $k \times k$ principal submatrices of $\hat{\Sigma}$ quickly becomes infeasible for even moderately large p and k .

In this paper, we address the question of whether it is possible to find an estimator of v_1 that is computable in (randomised) polynomial time, and that attains the minimax optimal rate of convergence when the sparsity of v_1 is allowed to vary with the sample size. Some progress in a related direction was made by [Berthet and Rigollet \(2013a,b\)](#), who considered the problem of testing the null hypothesis $H_0 : \Sigma = I_p$ against the alternative $H_1 : v^\top \Sigma v \geq 1 + \theta$ for some $v \in B_0(k)$ and $\theta > 0$. Of interest here is the minimal level $\theta = \theta_{n,p,k}$ that ensures small asymptotic testing error. Under a hypothesis on the computational intractability of a certain well-known problem from theoretical computer science (the ‘Planted Clique’ problem), Berthet and Rigollet showed that for certain classes of distributions, there is a gap between the minimal θ -level permitting successful detection with a randomised polynomial time test, and the corresponding θ -level when arbitrary tests are allowed.

The particular classes of distributions considered in [Berthet and Rigollet \(2013a,b\)](#) were highly tailored to the testing problem, and do not provide sufficient structure to study principal component estimation. The thesis of this paper, however, is that from the point

¹Here and below, $a_n \asymp b_n$ means $0 < \liminf_{n \rightarrow \infty} |a_n/b_n| \leq \limsup_{n \rightarrow \infty} |a_n/b_n| < \infty$.

²Since formal definitions of such notions from computational complexity theory may be unfamiliar to many statisticians, and to keep the paper as self-contained as possible, we provide a brief introduction to this topic in [Appendix D](#).

of view of both theory and applications, it is the estimation of sparse principal components, rather than testing for the existence of a distinguished direction, that is the more natural and fundamental (as well as more challenging) problem. It is by no means clear that the same phenomena should occur; in the example of k -SAT formulas, for instance, different results for statistical and computational trade-offs for estimation and testing were observed in [Feldman, Perkins and Vempala \(2013\)](#) and [Berthet \(2014\)](#) respectively.

Our first contribution, in [Section 2](#) is to introduce a new Restricted Covariance Concentration (RCC) condition that underpins the classes $\mathcal{P}_p(n, k, \theta)$ over which we perform the statistical and computational analyses (see [\(3\)](#) for a precise definition). The RCC condition is satisfied by subgaussian distributions, and moreover has the advantage of being more robust to certain mixture contaminations that turn out to be of key importance in the statistical analysis under the computational constraint. We show that subject to mild restrictions on the parameter values,

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}, v_1) \asymp \sqrt{\frac{k \log p}{n \theta^2}},$$

where $L(u, v) := \{1 - (u^\top v)^2\}^{1/2}$, and where no restrictions are placed on the class of estimators \hat{v} . By contrast, in [Section 3](#), we show that a variant \hat{v}^{SDP} of the semidefinite relaxation estimator of [d’Aspremont et al. \(2007\)](#) and [Bach, Ahipasaoglu and d’Aspremont \(2010\)](#), which is computable in polynomial time, satisfies

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}^{\text{SDP}}, v_1) \leq (16\sqrt{2} + 2) \sqrt{\frac{k^2 \log p}{n \theta^2}}.$$

Our main result, in [Section 4](#), is that under a weaker Planted Clique hypothesis than was assumed in [Berthet and Rigollet \(2013a,b\)](#), for any $\alpha \in (0, 1)$, there exists an asymptotic regime in which every sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators satisfies

$$\sqrt{\frac{n \theta^2}{k^{1+\alpha} \log p}} \sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}^{(n)}, v_1) \rightarrow \infty.$$

This result shows that there is a fundamental trade-off between statistical and computational efficiency in the estimation of sparse principal components, and that the estimator \hat{v}^{SDP} essentially achieves the optimal rate of convergence among estimators that are computable for even moderate datasets. Statistical and computational trade-offs have also recently been studied in the context of convex relaxation algorithms ([Chandrasekaran and Jordan](#),

2013), submatrix signal detection (Ma and Wu, 2013; Chen and Xu, 2014), sparse linear regression (Zhang, Wainwright and Jordan, 2014) and community detection (Hajek, Wu and Xu, 2014). Given the importance of computationally feasible algorithms with good statistical performance in today's era of Big Data, it seems clear that understanding the extent of this phenomenon in different settings will represent a key challenge for theoreticians in the coming years.

All proofs and several ancillary results are deferred to the Appendix. We end this section by introducing some notation used throughout the paper. For a vector $u = (u_1, \dots, u_M)^\top \in \mathbb{R}^M$, a matrix $A = (A_{ij}) \in \mathbb{R}^{M \times N}$ and for $q \in [1, \infty)$, we write $\|u\|_q := (\sum_{i=1}^M |u_i|^q)^{1/q}$ and $\|A\|_q := (\sum_{i=1}^M \sum_{j=1}^N |A_{ij}|^q)^{1/q}$ for their (entrywise) ℓ_q -norms. We also write $\|u\|_0 := \sum_{i=1}^M \mathbb{1}_{\{u_i \neq 0\}}$. For $S \subseteq \{1, \dots, M\}$ and $T \subseteq \{1, \dots, N\}$, we write $u_S := (u_i : i \in S)^\top$ and write $M_{S,T}$ for the $|S| \times |T|$ submatrix of M obtained by extracting the rows and columns with indices in S and T respectively.

2 Restricted Covariance Concentration and minimax rate of estimation

Let $p \geq 2$ and let \mathcal{P} denote the class of probability distributions P on \mathbb{R}^p with $\int_{\mathbb{R}^p} x dP(x) = 0$ and such that the entries of $\Sigma(P) := \int_{\mathbb{R}^p} xx^\top dP(x)$ are finite. For $P \in \mathcal{P}$, write $\lambda_1(P), \dots, \lambda_p(P)$ for the eigenvalues of $\Sigma(P)$, arranged in decreasing order. When $\lambda_1(P) - \lambda_2(P) > 0$, the first principal component $v_1(P)$, i.e. a unit-length eigenvector of Σ corresponding to the eigenvalue $\lambda_1(P)$, is well-defined up to sign. In some places below, and where it is clear from the context, we suppress the dependence of these quantities on P , or write the eigenvalues and eigenvectors as $\lambda_1(\Sigma), \dots, \lambda_p(\Sigma)$ and $v_1(\Sigma), \dots, v_p(\Sigma)$ respectively. Let X_1, \dots, X_n be independent and identically distributed random vectors with distribution P , and form the $n \times p$ matrix $\mathbf{X} := (X_1, \dots, X_n)^\top$. An *estimator* of v_1 is a measurable function from $\mathbb{R}^{n \times p}$ to \mathbb{R}^p , and we write $\mathcal{V}_{n,p}$ for the class of all such estimators.

Given unit vectors $u, v \in \mathbb{R}^p$, let $\Theta(u, v) := \cos^{-1}(|u^\top v|)$ denote the acute angle between u and v , and define the loss function

$$L(u, v) := \sin \Theta(u, v) = \{1 - (u^\top v)^2\}^{1/2} = \frac{1}{\sqrt{2}} \|uu^\top - vv^\top\|_2.$$

Note that $L(\cdot, \cdot)$ is invariant to sign changes of either of its arguments. The *directional variance* of P along a unit vector $u \in \mathbb{R}^p$ is defined as $V(u) := \mathbb{E}\{(u^\top X_1)^2\} = u^\top \Sigma u$. Its empirical counterpart is $\hat{V}(u) := n^{-1} \sum_{i=1}^n (u^\top X_i)^2 = u^\top \hat{\Sigma} u$, where $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$ denotes the sample covariance matrix.

Given $\ell \in \{1, \dots, p\}$ and $C \in (0, \infty)$, we say P satisfies a *Restricted Covariance Concentration* (RCC) condition with parameters p, n, ℓ and C , and write $P \in \text{RCC}_p(n, \ell, C)$, if

$$\mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq C \max\left(\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n}\right)\right\} \leq \delta \quad (2)$$

for all $\delta > 0$. It is also convenient to form classes $\text{RCC}_p(\ell, C) := \cap_{n=1}^\infty \text{RCC}_p(n, \ell, C)$ and $\text{RCC}_p(C) := \cap_{\ell=1}^p \text{RCC}_p(\ell, C)$. The RCC conditions amount to uniform Bernstein-type concentration properties of the directional variance around its expectation along all sparse directions. This condition turns out to be particularly convenient in the study of convergence rates in Sparse PCA, and moreover, as we show in Proposition 1 below, subgaussian distributions satisfy an RCC condition for all sample sizes n and all sparsity levels ℓ . Recall that a mean-zero distribution Q on \mathbb{R}^p is *subgaussian* with parameter³ $\sigma^2 \in (0, \infty)$, written $Q \in \text{subgaussian}_p(\sigma^2)$, if whenever $Y \sim Q$, we have $\mathbb{E}(e^{u^\top Y}) \leq e^{\sigma^2 \|u\|^2/2}$ for all $u \in \mathbb{R}^p$.

Proposition 1. (i) For every $\sigma > 0$, we have $\text{subgaussian}_p(\sigma^2) \subseteq \text{RCC}_p(16\sigma^2(1 + \frac{9}{\log p}))$.

(ii) In the special case $P = N_p(0, \Sigma)$, we have $P \in \text{RCC}_p(8\lambda_1(P)(1 + \frac{9}{\log p}))$.

Our convergence rate results for sparse principal component estimation will be proved over the following classes of distributions. For $\theta > 0$, let

$$\mathcal{P}_p(n, k, \theta) := \{P \in \text{RCC}_p(n, 2, 1) \cap \text{RCC}_p(n, 2k, 1) : v_1(P) \in B_0(k), \lambda_1(P) - \lambda_2(P) \geq \theta\}. \quad (3)$$

Observe that RCC classes have the scaling property that if the distribution of a random vector Y belongs to $\text{RCC}_p(n, \ell, C)$ and if $r > 0$, then the distribution of rY belongs to $\text{RCC}_p(n, \ell, r^2C)$. It is therefore convenient to fix $C = 1$ in both RCC classes in (3), so that θ becomes a measure of the signal-to-noise level.

For an arbitrary symmetric $p \times p$ matrix A , define $\hat{v}_{\max}^k(A) := \text{sargmax}_{u \in B_0(k)} u^\top A u$ to be the k -sparse maximum eigenvector of A , where sargmax denotes the smallest element of

³Note that some authors say that distributions satisfying this condition are subgaussian with parameter σ , rather than σ^2 .

the argmax in the lexicographic ordering. (This choice ensures that $\hat{v}_{\max}^k(A)$ is a measurable function of A .) Theorem 2 below gives a finite-sample minimax upper bound for estimating $v_1(P)$ over $\mathcal{P}_p(n, k, \theta)$. Similar bounds over Gaussian or subgaussian classes can be found in Cai, Ma and Wu (2013) and Vu and Lei (2013), who consider the more general problem of principal subspace estimation. As well as working with a larger class of distributions, our different proof techniques also facilitate an explicit constant.

Theorem 2. *For $2k \log p \leq n$, the k -sparse empirical maximum eigenvector, $\hat{v}_{\max}^k(\hat{\Sigma})$, satisfies*

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}_{\max}^k(\hat{\Sigma}), v_1(P)) \leq 2\sqrt{2} \left(1 + \frac{1}{\log p}\right) \sqrt{\frac{k \log p}{n\theta^2}} \leq 7\sqrt{\frac{k \log p}{n\theta^2}}. \quad (4)$$

A matching minimax lower bound of the same order in all parameters k, p, n and θ is given below. The proof techniques are adapted from Vu and Lei (2013).

Theorem 3. *Suppose that $7 \leq k \leq p^{1/2}$ and $0 < \theta \leq \frac{1}{16(1 + \frac{9}{\log p})}$. Then*

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}, v_1(P)) \geq \min \left\{ \frac{1}{1660} \sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}} \right\}.$$

We remark that the conditions in the statement of Theorem 3 can be strengthened or weakened, with a corresponding weakening or strengthening of the constants in the bound. For instance, a bound of the same order in k, p, n and θ could be obtained assuming only that $k \leq p^{1-\tau}$ for some $\tau > 0$. The upper bound on θ is also not particularly restrictive. For example, if $P = N_p(0, \sigma^2 I_p + \theta e_1 e_1^\top)$, where e_1 is the first standard basis vector in \mathbb{R}^p , then it can be shown that the condition $P \in \mathcal{P}_p(n, k, \theta)$ requires that $\theta \leq 1 - \sigma^2$.

3 Computationally efficient estimation

As was mentioned in the introduction, the trouble with the estimator $\hat{v}_{\max}^k(\hat{\Sigma})$ of Section 2, as well as the estimator of Cai, Ma and Wu (2013), is that there are no known polynomial time algorithms for their computation. In this section, we therefore study the (polynomial time) semidefinite relaxation estimator \hat{v}^{SDP} defined by the Algorithm 1 below. This estimator is a variant of one proposed by d'Aspremont et al. (2007), whose support recovery properties

were studied for a particular class of Gaussian distributions and a known sparsity level by [Amini and Wainwright \(2009\)](#).

To motivate the main step (Step 2) of Algorithm 1, it is convenient to let \mathcal{M} denote the class of $p \times p$ non-negative definite real, symmetric matrices, and let $\mathcal{M}_1 := \{M \in \mathcal{M} : \text{tr}(M) = 1\}$. Let $\mathcal{M}_{1,1}(k^2) := \{M \in \mathcal{M}_1 : \text{rank}(M) = 1, \|M\|_0 = k^2\}$ and observe that

$$\max_{u \in B_0(k)} u^\top \hat{\Sigma} u = \max_{u \in B_0(k)} \text{tr}(\hat{\Sigma} u u^\top) = \max_{M \in \mathcal{M}_{1,1}(k^2)} \text{tr}(\hat{\Sigma} M).$$

In the final expression, the rank and sparsity constraints are non-convex. We therefore adopt the standard semidefinite relaxation approach of dropping the rank constraint and replacing the sparsity (ℓ_0) constraint with an ℓ_1 penalty.

Algorithm 1: Pseudo-code for computing the semidefinite relaxation estimator \hat{v}^{SDP}

Input: $\mathbf{X} = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0, \epsilon > 0$

begin

Step 1: Set $\hat{\Sigma} \leftarrow n^{-1} \mathbf{X}^\top \mathbf{X}$.

Step 2: For $f(M) := \text{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let \hat{M} be an ϵ -maximiser of f in \mathcal{M}_1 . In other words, \hat{M} satisfies $f(\hat{M}) \geq \max_{M \in \mathcal{M}_1} f(M) - \epsilon$.

Step 3: Let $\hat{v}^{\text{SDP}} := \hat{v}_{\lambda, \epsilon}^{\text{SDP}} \leftarrow \text{sargmax}_{u: \|u\|_2=1} u^\top \hat{M} u$.

end

Output: \hat{v}^{SDP}

We now discuss the complexity of computing \hat{v}^{SDP} in detail. One possible way of implementing Step 2 is to use a generic interior-point method. However, as shown in [Nesterov \(2005\)](#), [Nemirovski \(2004\)](#) and [Bach, Ahipasaoglu and d'Aspremont \(2010\)](#), certain first-order algorithms (i.e. methods requiring $O(1/\epsilon)$ steps to find a feasible point achieving an ϵ -approximation of the optimal objective function value) can significantly outperform such generic interior-point solvers. The key idea in both [Nesterov \(2005\)](#) and [Nemirovski \(2004\)](#) is that the optimisation problem in Step 2 can be rewritten in a saddlepoint formulation:

$$\max_{M \in \mathcal{M}_1} \text{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 = \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \text{tr}((\hat{\Sigma} + U)M),$$

where $\mathcal{U} := \{U \in \mathbb{R}^{p \times p} : U^\top = U, \|U\|_\infty \leq \lambda\}$. The fact that $\text{tr}((\hat{\Sigma} + U)M)$ is linear in both M and U makes the problem amenable to proximal gradient methods. In Algorithm 2

below, we state a possible implementation of Step 2 of Algorithm 1, derived from the ‘basic implementation’ in Nemirovski (2004). In the algorithm, the $\|\cdot\|_2$ -norm projection $\Pi_{\mathcal{U}}(A)$ of a symmetric matrix $A = (A_{ij}) \in \mathbb{R}^{p \times p}$ onto \mathcal{U} is given by

$$(\Pi_{\mathcal{U}}(A))_{ij} := \text{sign}(A_{ij}) \min(|A_{ij}|, \lambda).$$

For the projection $\Pi_{\mathcal{M}_1}(A)$, first decompose $A =: PDP^\top$ for some orthogonal P and diagonal $D = \text{diag}(d)$, where $d = (d_1, \dots, d_p)^\top \in \mathbb{R}^p$. Now let $\Pi_{\mathcal{W}}(d)$ be the projection image of d on the unit $(p-1)$ -simplex $\mathcal{W} := \{(w_1, \dots, w_p) : w_j \geq 0, \sum_{j=1}^p w_j = 1\}$. Finally, transform back to obtain $\Pi_{\mathcal{M}_1}(A) := P \text{diag}(\Pi_{\mathcal{W}}(d)) P^\top$. The fact that Algorithm 2 outputs an ϵ -maximiser of the optimisation problem in Step 2 of Algorithm 1 follows from Nemirovski (2004, Theorem 3.2), which implies in our particular case that after N iterations,

$$\max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \text{tr}((\hat{\Sigma} + U)M) - \min_{U \in \mathcal{U}} \text{tr}((\hat{\Sigma} + U)\hat{M}) \leq \frac{\lambda^2 p^2 + 1}{\sqrt{2}N}.$$

Algorithm 2: A possible implementation of Step 2 of Algorithm 1

Input: $\hat{\Sigma} \in \mathcal{M}$, $\lambda > 0$, $\epsilon > 0$.

begin

Set $M_0 \leftarrow I_p/p$, $U_0 \leftarrow 0 \in \mathbb{R}^{p \times p}$ and $N \leftarrow \left\lceil \frac{\lambda^2 p^2 + 1}{\sqrt{2}\epsilon} \right\rceil$.

for $t \leftarrow 1$ **to** N **do**

$U'_t \leftarrow \Pi_{\mathcal{U}}(U_{t-1} - \frac{1}{\sqrt{2}}M_{t-1}), M'_t \leftarrow \Pi_{\mathcal{M}_1}(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}}U_{t-1}).$
 $U_t \leftarrow \Pi_{\mathcal{U}}(U_{t-1} - \frac{1}{\sqrt{2}}M'_t), M_t \leftarrow \Pi_{\mathcal{M}_1}(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}}U'_t).$

end

Set $\hat{M} \leftarrow \frac{1}{N} \sum_{t=1}^N M'_t$.

end

Output: \hat{M}

In Algorithm 1, Step 1 takes $O(np^2)$ floating point operations; Step 3 takes $O(p^3)$ operations in the worst case, though other methods such the Lanczos method (Lanczos, 1950; Golub and Van Loan, 1996) require only $O(p^2)$ operations under certain conditions. Our particular implementation (Algorithm 2) for Step 2 requires $O(\frac{\lambda^2 p^2 + 1}{\epsilon})$ iterations in the worst case, though this number may often be considerably reduced by terminating the **for** loop if the primal-dual gap

$$\lambda_1(\hat{U}_t + \hat{\Sigma}) - \{\text{tr}(\hat{M}_t \hat{\Sigma}) - \lambda \|\hat{M}_t\|_1\}$$

falls below ϵ , where $\hat{U}_t := t^{-1} \sum_{s=1}^t U'_s$ and $\hat{M}_t := t^{-1} \sum_{s=1}^t M'_s$. The most costly step within the **for** loop is the eigendecomposition used to compute the projection $\Pi_{\mathcal{M}_1}$, which takes $O(p^3)$ operations. Taking $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$ as in Theorem 5 below, we find an overall complexity for the algorithm of $O(\max(p^5, \frac{np^3}{\log p}))$ operations in the worst case.

We now turn to the theoretical properties of the estimator \hat{v}^{SDP} computed using Algorithm 1. Lemma 4 below is stated in a general, deterministic fashion, but will be used in Theorem 5 below to bound the loss incurred by the estimator on the event that the sample and population covariance matrices are close in ℓ_∞ -norm. See also Vu et al. (2013, Theorem 3.1) for a closely related result in the context of a projection matrix estimation problem.

Lemma 4. *Let $\Sigma \in \mathcal{M}$ be such that $\theta := \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\hat{\Sigma} := n^{-1} \mathbf{X}^\top \mathbf{X}$. For arbitrary $\lambda > 0$ and $\epsilon > 0$, if $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, then the semidefinite relaxation estimator \hat{v}^{SDP} in Algorithm 1 with inputs $\mathbf{X}, \lambda, \epsilon$ satisfies*

$$L(\hat{v}^{\text{SDP}}, v_1(\Sigma)) \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}}. \quad (5)$$

Theorem 5 below describes the statistical properties of the estimator \hat{v}^{SDP} over $\mathcal{P}_p(n, k, \theta)$ classes. It reveals in particular that we incur a loss of statistical efficiency of a factor of \sqrt{k} compared with the minimax upper bound in Theorem 2 in Section 2 above.

Theorem 5. *For an arbitrary $P \in \mathcal{P}_p(n, k, \theta)$ and $X_1, \dots, X_n \stackrel{iid}{\sim} P$, we write $\hat{v}^{\text{SDP}}(\mathbf{X})$ for the output of Algorithm 1 with input $\mathbf{X} := (X_1, \dots, X_n)^\top$, $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$. For any $k \in \{1, \dots, p\}$ and $n \in \mathbb{N}$ satisfying $4 \log p \leq n \leq k^2 p^2 \log p$, and for any $\theta \in (0, 1]$, we have*

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}^{\text{SDP}}(\mathbf{X}), v_1(P)) \leq (16\sqrt{2} + 2) \sqrt{\frac{k^2 \log p}{n\theta^2}}. \quad (6)$$

We remark that \hat{v}^{SDP} has the attractive property of being fully adaptive in the sense that it can be computed without knowledge of the sparsity level k . On the other hand, \hat{v}^{SDP} is not necessarily k -sparse. If a specific sparsity level is desired in a particular application, Algorithm 1 can be modified to obtain a (non-adaptive) k -sparse estimator having similar estimation risk. Specifically, we can find

$$\hat{v}_0^{\text{SDP}} \in \underset{u \in B_0(k)}{\operatorname{argmin}} L(\hat{v}^{\text{SDP}}, u).$$

Since $L(\hat{v}^{\text{SDP}}, u)^2 = 1 - (u^\top \hat{v}^{\text{SDP}})^2$, we can compute \hat{v}_0^{SDP} by setting all but the top k coordinates of \hat{v}^{SDP} in absolute value to zero and renormalising the vector. In particular, \hat{v}_0^{SDP} is computable in polynomial time. Furthermore, by the triangle inequality,

$$L(\hat{v}_0^{\text{SDP}}, v) \leq L(\hat{v}_0^{\text{SDP}}, \hat{v}^{\text{SDP}}) + L(\hat{v}^{\text{SDP}}, v) \leq 2L(\hat{v}^{\text{SDP}}, v).$$

We deduce that under the same conditions as in Theorem 5, and by a very similar argument,

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}_0^{\text{SDP}}, v_1(P)) \leq (32\sqrt{2} + 3) \sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

4 Computational lower bounds in sparse principal component estimation

Theorems 5 and 2 reveal a gap between the provable performance of our semidefinite relaxation estimator \hat{v}^{SDP} and the minimax optimal rate. It is natural to ask whether there exists a computationally efficient algorithm that achieves the statistically optimal rate of convergence. In fact, as we will see in Theorem 6 below, the rate of convergence given in (6) is essentially tight among the class of all *randomised polynomial time algorithms*⁴. Indeed, any randomised polynomial time algorithm with a faster rate of convergence could otherwise be adapted to solve instances of the Planted Clique problem that are believed to be hard; see Section 4.1 below for formal definitions and discussion. In this sense, the extra factor of \sqrt{k} is an intrinsic price in statistical efficiency that we have to pay for computational efficiency, and the estimator \hat{v}^{SDP} studied in Section 3 has essentially the best possible rate of convergence among computable estimators.

4.1 The Planted Clique problem

A *graph* $G := (V(G), E(G))$ is an ordered pair in which $V(G)$ is a countable set, and $E(G)$ is a subset of $\{\{x, y\} : x, y \in V(G), x \neq y\}$. For $x, y \in V(G)$, we say x and y are *adjacent*, and write $x \sim y$, if $\{x, y\} \in E(G)$. A *clique* C is a subset of $V(G)$ such that $\{x, y\} \in E(G)$ for all distinct $x, y \in C$. The problem of finding a clique of maximum size in a given graph

⁴In this section, terms from computational complexity theory defined in Appendix D are written in italics at their first occurrence.

G is known to be NP-complete (Karp, 1972). It is therefore natural to consider randomly generated input graphs with a clique ‘planted’ in, where the signal is much less confounded by the noise. Such problems were first suggested by Jerrum (1992) and Kučera (1995) as a potentially easier variant of the classical Clique problem.

Let \mathbb{G}_m denote the collection of all graphs with m vertices. Define \mathcal{G}_m to be the distribution on \mathbb{G}_m associated with the standard Erdős–Rényi random graph. In other words, under \mathcal{G}_m , each pair of vertices is adjacent independently with probability $1/2$. For any $\kappa \in \{1, \dots, m\}$, let $\mathcal{G}_{m,\kappa}$ be a distribution on \mathbb{G}_m constructed by first picking κ distinct vertices uniformly at random and connecting all edges (the ‘planted clique’), then joining each remaining pair of distinct vertices by an edge independently with probability $1/2$. The Planted Clique problem has input graphs randomly sampled from the distribution $\mathcal{G}_{m,\kappa}$. Due to the random nature of the problem, the goal of the Planted Clique problem is to find (possibly randomised) algorithms that can locate a maximum clique K_m with high probability.

It is well known that, for a standard Erdős–Rényi graph, $\frac{|K_m|}{2^{\log_2 m}} \xrightarrow{\text{a.s.}} 1$ (e.g. Grimmett and McDiarmid, 1975). If $\kappa = \kappa_m$ is such that $\liminf_{m \rightarrow \infty} \frac{\kappa}{2^{\log_2 m}} > 1$, it can in fact be shown that the planted clique is asymptotically almost surely also the unique maximum clique in the input graph. As pointed out in Kučera (1995), there exists $C > 0$ such that, if $\kappa > C\sqrt{m \log m}$, then asymptotically almost surely, vertices in the planted clique have larger degrees than all other vertices, in which case they can be located in $O(m^2)$ operations. Alon, Krivelevich and Sudakov (1998) improved the above result by exhibiting a spectral method that, given any $c > 0$, identifies planted cliques of size $\kappa \geq c\sqrt{m}$ asymptotically almost surely.

Although several other polynomial time algorithms have subsequently been discovered for the $\kappa \geq c\sqrt{m}$ case (e.g. Feige and Krauthgamer, 2000; Feige and Ron, 2010; Ames and Vavasis, 2011), there is no known randomised polynomial time algorithm that can detect below this threshold. Jerrum (1992) hinted at the hardness of this problem by showing that a specific Markov chain approach fails to work when $\kappa = O(m^{1/2-\delta})$ for some $\delta > 0$. Feige and Krauthgamer (2003) showed that Lovász–Schrijver semidefinite programming relaxation methods also fail in this regime. Feldman et al. (2013) recently presented further evidence of the hardness of this problem by showing that a broad class of algorithms, which they refer to as ‘statistical algorithms’, cannot solve the Planted Clique problem with $\kappa = O(m^{1/2-\delta})$.

in randomised polynomial time, for any $\delta > 0$. It is now widely accepted in theoretical computer science that the Planted Clique problem is hard, in the sense that the following assumption holds:

(A1) For any sequence $\kappa = \kappa_m$ such that $\kappa \leq m^\beta$ for some $0 < \beta < 1/2$, there is no randomised polynomial time algorithm that can correctly identify the planted clique with probability tending to 1 as $m \rightarrow \infty$.

Researchers have used the hardness of the planted clique problem as an assumption to prove various impossibility results in other problems. Examples include cryptographic applications (Juels and Peinado, 2000; Applebaum, Barak and Wigderson, 2010), testing k -wise independence (Alon et al., 2007) and approximating Nash equilibria (Hazan and Krauthgamer, 2011). Recent works by Berthet and Rigollet (2013a,b) and Ma and Wu (2013) used similar hypotheses on the hardness of the Planted Clique problem to establish computational lower bounds in sparse principal component detection and sparse submatrix detection problems respectively. It is worth noting that our Assumption **(A1)** is weak in the sense that it is implied by Hypothesis A_{PC} of Berthet and Rigollet (2013b) and Hypothesis 1 of Ma and Wu (2013).

4.2 Computational lower bounds

In this section, we use a reduction argument to show that, under Assumption **(A1)**, it is impossible to achieve the statistically optimal rate of sparse principal component estimation using randomised polynomial time algorithms. For $\rho \in \mathbb{N}$, and for $x \in \mathbb{R}$, we let $[x]_\rho$ denote x in its binary representation, rounded to ρ significant figures. Let $[\mathbb{R}]_\rho := \{[x]_\rho : x \in \mathbb{R}\}$. We say $(\hat{v}^{(n)})$ is a *sequence of randomised polynomial time estimators* of $v_1 \in \mathbb{R}^{p_n}$ if $\hat{v}^{(n)}$ is a measurable function from $\mathbb{R}^{n \times p_n}$ to \mathbb{R}^{p_n} and if, for every $\rho \in \mathbb{N}$, there exists a randomised polynomial time algorithm M_{pr} such that for any $\mathbf{x} \in ([\mathbb{R}]_\rho)^{n \times p_n}$ we have $[\hat{v}^{(n)}(\mathbf{x})]_\rho = [M_{pr}(\mathbf{x})]_\rho$. The sequence of semidefinite programming estimators (\hat{v}^{SDP}) defined in Section 3 is an example of a sequence of randomised polynomial estimators of $v_1(P)$.

Theorem 6. Assume **(A1)**, and let $\alpha \in (0, 1)$. For any $n \in \mathbb{N}$, let $k = k_n := \lfloor n^{2/(5-\alpha)} \rfloor$, $p = p_n := n$ and $\theta = \theta_n := n^{-(1-\alpha)/(5-\alpha)}/1000$. For $P \in \mathcal{P}_p(n, k, \theta)$, let \mathbf{X} be an $n \times p$ matrix with independent rows, each having distribution P . Then every sequence $(\hat{v}^{(n)})$ of

randomised polynomial time estimators of $v_1(P)$ satisfies

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha}\log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{(n)}(\mathbf{X}), v_1(P)) \rightarrow \infty \quad (7)$$

as $n \rightarrow \infty$.

In the statement of the theorem, we chose a particular sequence of parameters (k_n, p_n, θ_n) in order to present the main idea of our result without minimal notational clutter. However, it can be shown that the result described in Theorem 6 holds true for a range of values of the parameters. For instance, if the sequence (k_n, p_n, θ_n) satisfies $nk_n^{3-\alpha}/p_n^2 \rightarrow \infty$, $k_n = O(p_n^{1/2-\delta})$ for some $\delta > 0$ and $\theta_n = ck^2/p_n$ for sufficiently small $c > 0$, then the conclusion of Theorem 6 still holds. We also remark that, in cases where the signal-to-noise ratio is very high, it is possible to attain the minimax rate of convergence with a polynomial time algorithm (Ma, 2013; Wang, Lu and Liu, 2014), so the statistical and computational trade-off disappears in such settings.

The proof of Theorem 6 relies on a randomised polynomial time reduction from the Planted Clique problem to the sparse principal component estimation problem. The reduction is adapted from the ‘bottom-left transformation’ of Berthet and Rigollet (2013b), and requires a rather different and delicate analysis. We sketch the main idea below.

Given any graph $G \sim \mathcal{G}_{m,\kappa}$, we generate an $n \times p$ off-diagonal submatrix \mathbf{A} uniformly at random from the adjacency matrix of G . We then replace each 0 with -1 in \mathbf{A} and flip the signs of each row of \mathbf{A} independently with probability $1/2$ to obtain a new matrix \mathbf{X} . When n and p are not too large compared to m , the joint distribution of the set of row vectors of \mathbf{X} will be very close in total variation distance to a mixture of joint distributions of a set of independent and identically distributed vectors. Moreover, after scaling, and for a suitable range of n, p, k and θ , each component of the mixture can be expressed as $P^{\otimes n}$ for some $P \in \mathcal{P}_p(n, k, \theta)$. We then show that if the conclusion of the theorem were false, so that there would exist a randomised polynomial time algorithm for estimating the sparse principal component of P with rate of convergence $O(\sqrt{\frac{k^{1+\alpha}\log p}{n\theta^2}})$, then the largest k absolute entries \hat{S} of this estimator would index at least $3k/4$ planted clique vertices with high probability. The remainder of the planted clique vertices could then be identified, also with high probability, by adding vertices connected to a large number of vertices indexed in \hat{S} .

Acknowledgements

The first author is supported by a Benefactors' scholarship from St John's College, Cambridge. The second author is supported by Air Force of Scientific Research (AFOSR) grant FA9550-14-1-0098 at the Center for the Mathematics of Information at the California Institute of Technology. The third author is supported by Engineering and Physical Sciences Research Council Early Career Fellowship EP/J017213/1.

A Appendix: Proofs from Sections 2 and 3

Proof of Proposition 1. (i) Let $P \in \text{subgaussian}_p(\sigma^2)$, and let $X_1, \dots, X_n \stackrel{iid}{\sim} P$. Then, for any $u \in B_0(\ell)$ and $t \geq 0$, we have

$$\mathbb{P}(u^\top X_1 \geq t) \leq e^{-t^2/\sigma^2} \mathbb{E}(e^{tu^\top X_1/\sigma^2}) \leq e^{-t^2/(2\sigma^2)}.$$

Similarly, $\mathbb{P}(-u^\top X_1 \geq t) \leq e^{-t^2/(2\sigma^2)}$. Write $\mu_u := \mathbb{E}\{(u^\top X_1)^2\}$; since

$$1 + \frac{1}{2}\mu_u t^2 + o(t^2) = \mathbb{E}(e^{tu^\top X_1}) \leq e^{t^2\sigma^2/2} = 1 + \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

as $t \rightarrow 0$, we deduce that $\mu_u \leq \sigma^2$. Now, for any integer $m \geq 2$,

$$\begin{aligned} & \mathbb{E}(|(u^\top X_1)^2 - \mu_u|^m) \\ &= \mathbb{E}(\{(u^\top X_1)^2 - \mu_u\}^m \mathbb{1}_{\{(u^\top X_1)^2 \geq \mu_u\}}) + \mathbb{E}(\{\mu_u - (u^\top X_1)^2\}^m \mathbb{1}_{\{\mu_u > (u^\top X_1)^2\}}) \\ &\leq \int_0^\infty \mathbb{P}\left[\{(u^\top X_1)^2 - \mu_u\}^m \mathbb{1}_{\{(u^\top X_1)^2 \geq \mu_u\}} \geq t\right] dt + \mu_u^m \\ &\leq 2 \int_0^\infty e^{-\frac{t^{1/m} + \mu_u}{2\sigma^2}} dt + \mu_u^m = m!(2\sigma^2)^m \left\{ 2e^{-\mu_u/(2\sigma^2)} + \frac{1}{m!} \left(\frac{\mu_u}{2\sigma^2}\right)^m \right\} \leq 2m!(2\sigma^2)^m, \end{aligned}$$

where the final inequality follows because the function $x \mapsto 2e^{-x} + x^m/m!$ is decreasing on $[0, 1/2]$. This calculation allows us to apply Bernstein's inequality (e.g. [van de Geer, 2000](#), Lemma 5.7, taking $K = 2\sigma^2, R = 4\sigma^2$ in her notation), to deduce that for any $s \geq 0$,

$$\mathbb{P}(|\hat{V}(u) - V(u)| \geq s) \leq 2 \exp\left(-\frac{ns^2}{4\sigma^2 s + 32\sigma^4}\right).$$

It follows by Lemma 9 in Appendix C, taking $\epsilon = 1/4$ in that result, that if $\eta > 0$ is such that $\ell \log(p/\eta) \leq n$, then for $C := 8\sigma^2$, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 2C\sqrt{\frac{\ell \log(p/\eta)}{n}}\right) \\ \leq 2\pi\ell^{1/2} \binom{p}{\ell} \left(\frac{128}{\sqrt{255}}\right)^{\ell-1} \exp\left(-\frac{C^2\ell \log(p/\eta)}{4C\sigma^2\sqrt{\frac{\ell \log(p/\eta)}{n}} + 32\sigma^4}\right) \\ \leq 2\pi\ell^{1/2} \left(\frac{e}{\ell}\right)^\ell \left(\frac{128}{\sqrt{255}}\right)^{\ell-1} \eta^\ell \leq e^9\eta, \end{aligned}$$

Similarly, if $\ell \log(p/\eta) > n$, then

$$\begin{aligned} \mathbb{P}\left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 2C\frac{\ell \log(p/\eta)}{n}\right) \\ \leq 2\pi\ell^{1/2} \binom{p}{\ell} \left(\frac{128}{\sqrt{255}}\right)^{\ell-1} \exp\left(-\frac{C^2\ell^2 \log^2(p/\eta)}{4C\sigma^2\ell \log(p/\eta) + 32\sigma^4 n}\right) \leq e^9\eta. \end{aligned}$$

Setting $\delta := e^9\eta$, we find (noting that we only need to consider the case $\delta \in (0, 1]$) that

$$\begin{aligned} \mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2\left(1 + \frac{9}{\log p}\right) \max\left(\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n}\right)\right\} \\ \leq \mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2 \max\left(\sqrt{\frac{\ell \log(e^9 p/\delta)}{n}}, \frac{\ell \log(e^9 p/\delta)}{n}\right)\right\} \leq \delta. \end{aligned}$$

(ii) An immediate consequence of Lemma 1 of [Laurent and Massart \(2000\)](#) is that if Y_1, \dots, Y_n are independent χ_1^2 random variables, then we have for all $a > 0$ that

$$\mathbb{P}\left(\frac{1}{n} \left|\sum_{i=1}^n Y_i - 1\right| \geq a\right) \leq 2e^{-\frac{n}{2}(1+a-\sqrt{1+2a})} \leq 2e^{-n \min(\frac{a}{4}, \frac{a^2}{16})}.$$

Setting $\eta := e^{-n \min(\frac{a}{4}, \frac{a^2}{16})}$, we deduce that

$$\mathbb{P}\left\{\frac{1}{n} \left|\sum_{i=1}^n Y_i - 1\right| \geq 4 \max\left(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\right)\right\} \leq 2\eta.$$

Hence, using Lemma 9 again, and by a similar calculation to Part (i),

$$\mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 8\lambda_1(P) \max\left(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\right)\right\} \leq e^9 p^\ell \eta.$$

The result follows on setting $\delta := e^9 p^\ell \eta$. □

Proof of Theorem 2. Fix an arbitrary $P \in \mathcal{P}_p(n, k, \theta)$. For notational simplicity, we write $v := v_1(P)$ and $\hat{v} := \hat{v}_{\max}^k(\hat{\Sigma})$ in this proof. We now exploit the Curvature Lemma of [Vu et al. \(2013, Lemma 3.1\)](#), which is closely related to the Davis–Kahan $\sin \theta$ theorem ([Davis and Kahan, 1970](#); [Yu, Wang and Samworth, 2014](#)). This lemma gives that

$$\|\hat{v}\hat{v}^\top - vv^\top\|_2^2 \leq \frac{2}{\theta} \text{tr}(\Sigma(vv^\top - \hat{v}\hat{v}^\top)) \leq \frac{2}{\theta} \text{tr}((\Sigma - \hat{\Sigma})(vv^\top - \hat{v}\hat{v}^\top)).$$

When $\hat{v}\hat{v}^\top \neq vv^\top$, we have that $\frac{vv^\top - \hat{v}\hat{v}^\top}{\|vv^\top - \hat{v}\hat{v}^\top\|_2}$ has rank 2, trace 0 and has non-zero entries in at most $2k$ rows and $2k$ columns. It follows that its non-zero eigenvalues are $\pm 1/\sqrt{2}$, so it can be written as $(xx^\top - yy^\top)/\sqrt{2}$ for some $x, y \in B_0(2k)$. Thus

$$\begin{aligned} \mathbb{E}_P L(\hat{v}, v) &= \mathbb{E}_P \frac{1}{\sqrt{2}} \|\hat{v}\hat{v}^\top - vv^\top\|_2 \leq \frac{1}{\theta} \mathbb{E}_P \text{tr}((\Sigma - \hat{\Sigma})(xx^\top - yy^\top)) \\ &\leq \frac{2}{\theta} \mathbb{E}_P \sup_{u \in B_0(2k)} |\hat{V}(u) - V(u)| \leq 2\sqrt{2} \left(1 + \frac{1}{\log p}\right) \sqrt{\frac{k \log p}{n\theta^2}}, \end{aligned}$$

where we have used Proposition 8 in Appendix C to obtain the final inequality. \square

Proof of Theorem 3. Set $\sigma^2 := \frac{1}{8(1+\frac{9}{\log p})} - \theta$. We have by Proposition 1(ii) that $N_p(0, \sigma^2 I_p + \theta v_1 v_1^\top) \in \mathcal{P}_p(n, k, \theta)$ for any unit vector $v_1 \in B_0(k)$. Define $k_0 := k - 1$ and $p_0 := p - 1$. Applying the variant of the Gilbert–Varshamov lemma given as Lemma 10 in Appendix C with $\alpha := 1/2$ and $\beta := 1/4$, we can construct a set \mathcal{N}_0 of k_0 -sparse vectors in $\{0, 1\}^{p_0}$ with cardinality at least $(p_0/k_0)^{k_0/8}$, such that the Hamming distance between every pair of distinct points in \mathcal{N}_0 is at least k_0 . For $\epsilon \in (0, 1]$ to be chosen later, define a set of k -sparse vectors in \mathbb{R}^p by

$$\mathcal{N} := \left\{ \begin{pmatrix} \sqrt{1 - \epsilon^2} \\ k_0^{-1/2} \epsilon u_0 \end{pmatrix} : u_0 \in \mathcal{N}_0 \right\}.$$

Observe that if u, v are distinct elements of \mathcal{N} , then

$$L(u, v) = \{1 - (u^\top v)^2\}^{1/2} \geq \{1 - (1 - \epsilon^2/2)^2\}^{1/2} \geq \frac{\sqrt{3}\epsilon}{2},$$

and similarly $L(u, v) \leq \epsilon$. For $u \in \mathcal{N}$, let P_u denote the $N_p(0, \sigma^2 I_p + \theta uu^\top)$ distribution. For any estimator $\hat{v} \in \mathcal{V}_{n,p}$, we define $\hat{\psi}_{\hat{v}} := \text{sargmin}_{u \in \mathcal{N}} L(\hat{v}, u)$, where sargmin denotes the smallest element of the argmin in the lexicographic ordering. Note that $\{\hat{\psi}_{\hat{v}} \neq u\} \subseteq \{L(\hat{v}, u) \geq \sqrt{3}\epsilon/4\}$. We now apply the generalised version of Fano’s lemma given as Lemma 11 in Appendix C. Writing $D(P\|Q)$ for the Kullback–Leibler divergence between two probability

measures defined on the same space (a formal definition is given just prior to Lemma 11), we have

$$\begin{aligned}
\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1(P)) &\geq \inf_{\hat{v} \in \mathcal{V}_{n,p}} \max_{u \in \mathcal{N}} \mathbb{E}_{P_u} L(\hat{v}, u) \\
&\geq \frac{\sqrt{3}\epsilon}{4} \inf_{\hat{v} \in \mathcal{V}_{n,p}} \max_{u \in \mathcal{N}} P_u^{\otimes n}(\hat{\psi}_{\hat{v}} \neq u) \\
&\geq \frac{\sqrt{3}\epsilon}{4} \left(1 - \frac{\max_{u,v \in \mathcal{N}, u \neq v} D(P_v^{\otimes n} \| P_u^{\otimes n}) + \log 2}{(k_0/8) \log(p_0/k_0)} \right). \tag{8}
\end{aligned}$$

We can compute, for distinct points $u, v \in \mathcal{N}$,

$$\begin{aligned}
D(P_v^{\otimes n} \| P_u^{\otimes n}) &= nD(P_v \| P_u) = \frac{n}{2} \text{tr}((\sigma^2 I_p + \theta uu^\top)^{-1}(\sigma^2 I_p + \theta vv^\top) - I_p) \\
&= \frac{n}{2} \text{tr}((\sigma^2 I_p + \theta uu^\top)^{-1} \theta(vv^\top - uu^\top)) \\
&= \frac{n\theta}{2} \text{tr} \left(\left(\frac{1}{\sigma^2} I_p - \frac{\theta}{\sigma^2(\sigma^2 + \theta)} uu^\top \right) (vv^\top - uu^\top) \right) \\
&= \frac{n\theta^2}{2\sigma^2(\sigma^2 + \theta)} L^2(u, v) \leq \frac{n\theta^2 \epsilon^2}{2\sigma^2(\sigma^2 + \theta)}. \tag{9}
\end{aligned}$$

Let

$$a := 1 - \frac{8 \log 2}{k_0 \log(p_0/k_0)}, \quad b := \frac{4n\theta^2}{\sigma^2(\sigma^2 + \theta)k_0 \log(p_0/k_0)} \quad \text{and} \quad \epsilon := \min \left\{ \sqrt{\frac{a}{3b}}, 1 \right\}.$$

Then from (8) and (9), we find that

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1(P)) \geq \min \left\{ \frac{1}{1660} \sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}} \right\},$$

as required. \square

Proof of Lemma 4. For convenience, we write $v := v_1(\Sigma)$ and \hat{v} for \hat{v}^{SDP} in this proof. We first study $vv^\top - \hat{M}$, where $\hat{M} \in \mathcal{M}_1$ is computed in Step 2 of Algorithm 1. By the Curvature Lemma of Vu et al. (2013, Lemma 3.1),

$$\|vv^\top - \hat{M}\|_2^2 \leq \frac{2}{\theta} \text{tr}(\Sigma(vv^\top - \hat{M})).$$

Moreover, since $vv^\top \in \mathcal{M}_1$, we have the basic inequality

$$\text{tr}(\hat{\Sigma}\hat{M}) - \lambda \|\hat{M}\|_1 \geq \text{tr}(\hat{\Sigma}vv^\top) - \lambda \|vv^\top\|_1 - \epsilon.$$

Let S denote the set of indices corresponding to the non-zero components of v , and recall that $|S| \leq k$. Then, since by hypothesis $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, we have

$$\begin{aligned} \|vv^\top - \hat{M}\|_2^2 &\leq \frac{2}{\theta} \left\{ \text{tr}(\hat{\Sigma}(vv^\top - \hat{M})) + \text{tr}((\Sigma - \hat{\Sigma})(vv^\top - \hat{M})) \right\} \\ &\leq \frac{2}{\theta} (\lambda \|vv^\top\|_1 - \lambda \|\hat{M}\|_1 + \epsilon + \|\hat{\Sigma} - \Sigma\|_\infty \|vv^\top - \hat{M}\|_1) \\ &\leq \frac{2\lambda}{\theta} (\|v_S v_S^\top\|_1 - \|\hat{M}_{S,S}\|_1 + \|v_S v_S^\top - \hat{M}_{S,S}\|_1) + \frac{2\epsilon}{\theta} \\ &\leq \frac{4\lambda}{\theta} \|v_S v_S^\top - \hat{M}_{S,S}\|_1 + \frac{2\epsilon}{\theta} \leq \frac{4\lambda k}{\theta} \|vv^\top - \hat{M}\|_2 + \frac{2\epsilon}{\theta}. \end{aligned}$$

We deduce that

$$\|vv^\top - \hat{M}\|_2 \leq \frac{4\lambda k}{\theta} + \sqrt{\frac{2\epsilon}{\theta}}.$$

On the other hand,

$$\begin{aligned} \|vv^\top - \hat{M}\|_2^2 &= \text{tr}((vv^\top - \hat{M})^2) = 1 - 2v^\top \hat{M} v + \text{tr}(\hat{M}^2) \\ &\geq 1 - 2\hat{v}^\top \hat{M} \hat{v} + \text{tr}(\hat{M}^2) = \|\hat{v}\hat{v}^\top - \hat{M}\|_2^2. \end{aligned}$$

We conclude that

$$\begin{aligned} L(\hat{v}, v) &= \frac{1}{\sqrt{2}} \|\hat{v}\hat{v}^\top - vv^\top\|_2 \leq \frac{1}{\sqrt{2}} (\|\hat{v}\hat{v}^\top - \hat{M}\|_2 + \|vv^\top - \hat{M}\|_2) \leq \sqrt{2} \|vv^\top - \hat{M}\|_2 \\ &\leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}}, \end{aligned}$$

as required. \square

Proof of Theorem 5. By Lemma 4, as well as Lemma 12 in Appendix C,

$$\begin{aligned} \mathbb{E}_P L(\hat{v}^{\text{SDP}}, v_1(P)) &\leq \mathbb{E}_P \left\{ L(\hat{v}^{\text{SDP}}, v_1(P)) \mathbb{1}_{\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda\}} \right\} + \mathbb{E}_P \left\{ L(\hat{v}^{\text{SDP}}, v_1(P)) \mathbb{1}_{\{\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\}} \right\} \\ &\leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}} + \mathbb{P} \left(\sup_{u \in B_0(2)} |\hat{V}(u) - V(u)| > 2\sqrt{\frac{\log p}{n}} \right) \end{aligned} \quad (10)$$

Since $P \in \text{RCC}_p(n, 2, 1)$,

$$\mathbb{P} \left\{ \sup_{u \in B_0(2)} |\hat{V}(u) - V(u)| > \max \left(\sqrt{\frac{2 \log(p/\delta)}{n}}, \frac{2 \log(p/\delta)}{n} \right) \right\} \leq \delta.$$

Set $\delta := \sqrt{\frac{k^2 \log p}{n}}$. Under the assumptions of the theorem, we have that $\log(1/\delta) \leq \log p$, which implies that $\sqrt{\frac{2 \log(p/\delta)}{n}} \leq \sqrt{\frac{4 \log p}{n}} \leq 1$. Consequently,

$$\mathbb{P} \left(\sup_{u \in B_0(2)} |\hat{V}(u) - V(u)| > 2\sqrt{\frac{\log p}{n}} \right) \leq \sqrt{\frac{k^2 \log p}{n}}. \quad (11)$$

The desired risk bound follows from (10) and (11). \square

B Proof of Theorem 6

Proof of Theorem 6. Suppose, for a contradiction, that there exist an infinite subset \mathcal{N} of \mathbb{N} , $K_0 \in [0, \infty)$ and a sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators of $v_1(P)$ satisfying

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L(\hat{v}^{(n)}(\mathbf{X}), v_1(P)) \leq K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

for all $n \in \mathcal{N}$. Let $L := \lceil \log n \rceil$, let $m = m_n := \lceil 10Lp_n/9 \rceil$ and let $\kappa = \kappa_n := Lk_n$. We claim that Algorithm 3 below is a randomised polynomial time algorithm that correctly identifies the Planted Clique problem on m_n vertices and a planted clique of size κ_n with probability tending to 1 as $n \rightarrow \infty$. Since $\kappa_n = O(m_n^{2/(5-\alpha)} \log m_n)$, this contradicts Assumption (A1). We prove the claim below.

Algorithm 3: Pseudo-code for a planted clique algorithm based on a hypothetical randomised polynomial time sparse principal component estimation algorithm.

Input: $m \in \mathbb{N}$, $\kappa \in \{1, \dots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$

begin

Step 1: Let $n \leftarrow \lfloor 9m/(10L) \rfloor$, $p \leftarrow n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \dots, u_n, w_1, \dots, w_p$ uniformly at random without replacement from $V(G)$. Form

$\mathbf{A} = (A_{ij}) \leftarrow (\mathbb{1}_{\{u_i \sim w_j\}}) \in \mathbb{R}^{n \times p}$ and $\mathbf{X} \leftarrow \text{diag}(\xi_1, \dots, \xi_n)(2\mathbf{A} - \mathbf{1}_{n \times p})$, where ξ_1, \dots, ξ_n are independent Rademacher random variables (independent of $u_1, \dots, u_n, w_1, \dots, w_p$), and where every entry of $\mathbf{1}_{n \times p} \in \mathbb{R}^{n \times p}$ is 1.

Step 2: Use the randomised estimator $\hat{v}^{(n)}$ to compute $\hat{v} = \hat{v}^{(n)}(\mathbf{X}/\sqrt{750})$.

Step 3: Let $\hat{S} = \hat{S}(\hat{v})$ be the lexicographically smallest k -subset of $\{1, \dots, p\}$ such that $(\hat{v}_j : j \in \hat{S})$ contains the k largest coordinates of \hat{v} in absolute value.

Step 4: For $u \in V(G)$ and $W \subseteq V(G)$, let $\text{nb}(u, W) := \mathbb{1}_{\{u \in W\}} + \sum_{w \in W} \mathbb{1}_{\{u \sim w\}}$. Set $\hat{K} := \{u \in V(G) : \text{nb}(u, \{w_j : j \in \hat{S}\}) \geq 3k/4\}$.

end

Output: \hat{K}

Let $G \sim \mathbb{G}_{m, \kappa}$, and let $K \subseteq V(G)$ denote the planted clique. Note that the matrix \mathbf{A} defined in Step 1 of Algorithm 3 is the off-diagonal block of the adjacency matrix of G associated with the bipartite graph induced by the two parts $\{u_i : i = 1, \dots, n\}$ and

$\{w_j : j = 1, \dots, p\}$. Let $\boldsymbol{\epsilon}' = (\epsilon'_1, \dots, \epsilon'_n)^\top$ and $\boldsymbol{\gamma}' = (\gamma'_1, \dots, \gamma'_p)^\top$, where $\epsilon'_i := \mathbb{1}_{\{u_i \in K\}}$, $\gamma'_j := \mathbb{1}_{\{w_j \in K\}}$, and set $S' := \{j : \gamma'_j = 1\}$.

It is convenient at this point to introduce the notion of a *Graph Vector distribution*. We say Y has a p -variate Graph Vector distribution with parameters $g = (g_1, \dots, g_p)^\top \in \{0, 1\}^p$ and $\pi_0 \in [0, 1]$, and write $Y \sim \text{GV}_p^g(\pi_0)$, if we can write

$$Y = \xi \{(1 - \epsilon)R + \epsilon(g + \tilde{R})\},$$

where ξ , ϵ and R are independent, where ξ is a Rademacher random variable, where $\epsilon \sim \text{Bern}(\pi_0)$, where $R = (R_1, \dots, R_p)^\top \in \mathbb{R}^p$ has independent Rademacher components, and where $\tilde{R} = (\tilde{R}_1, \dots, \tilde{R}_p)^\top$ with $\tilde{R}_j := (1 - g_j)R_j$.

Let $(\boldsymbol{\epsilon}, \boldsymbol{\gamma})^\top = (\epsilon_1, \dots, \epsilon_n, \gamma_1, \dots, \gamma_p)^\top$ be $n + p$ independent $\text{Bern}(\kappa/m)$ random variables. For $i = 1, \dots, n$, let $Y_i := \xi_i \{(1 - \epsilon_i)R_i + \epsilon_i(\gamma + \tilde{R}_i)\}$ so that, conditional on $\boldsymbol{\gamma}$, the random vectors Y_1, \dots, Y_n are independent, each distributed as $\text{GV}_p^\gamma(\kappa/m)$. As shorthand, we denote this conditional distribution as Q_γ , and write $S := \{j : \gamma_j = 1\}$. Note that by Lemma 7 below, $Q_\gamma \in \cap_{\ell=1}^{\lfloor 20p/(9k) \rfloor} \text{RCC}_p(\ell, 750)$.

Let $\mathbf{Y} := (Y_1, \dots, Y_n)^\top$. Recall that if P and Q are probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, the *total variation distance* between P and Q is defined by

$$d_{\text{TV}}(P, Q) := \sup_{B \in \mathcal{B}} |P(B) - Q(B)|.$$

Writing $\mathcal{L}(Z)$ for the distribution (or law) of a generic random element Z , and using elementary properties of the total variation distance given in Lemma 14 in Appendix C, we have

$$\begin{aligned} d_{\text{TV}}(\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y})) &\leq d_{\text{TV}}\left(\mathcal{L}(\boldsymbol{\epsilon}', \boldsymbol{\gamma}', (R_{ij}), (\xi_i)), \mathcal{L}(\boldsymbol{\epsilon}, \boldsymbol{\gamma}, (R_{ij}), (\xi_i))\right) \\ &= d_{\text{TV}}(\mathcal{L}(\boldsymbol{\epsilon}', \boldsymbol{\gamma}'), \mathcal{L}(\boldsymbol{\epsilon}, \boldsymbol{\gamma})) \leq \frac{2(n+p)}{m} \leq \frac{18}{5L}. \end{aligned} \quad (12)$$

Here, the penultimate inequality follows from Diaconis and Freedman (1980, Theorem 4). In view of (12), we initially analyse Steps 2, 3 and 4 in Algorithm 3 with \mathbf{X} replaced by \mathbf{Y} . Observe that $\mathbb{E}(Y_i | \boldsymbol{\gamma}) = 0$ and, writing $\Delta := \text{diag}(\boldsymbol{\gamma}) \in \mathbb{R}^{p \times p}$, we have

$$\Sigma_\gamma := \text{Cov}(Y_i | \boldsymbol{\gamma}) = \mathbb{E}\{(1 - \epsilon_i)R_i R_i^\top + \epsilon_i(\gamma + \tilde{R}_i)(\gamma + \tilde{R}_i)^\top | \boldsymbol{\gamma}\} = I_p + \frac{\kappa}{m}(\boldsymbol{\gamma} \boldsymbol{\gamma}^\top - \Delta).$$

Writing $N_\gamma := \sum_{j=1}^p \gamma_j$, it follows that the largest eigenvalue of Σ_γ is $1 + \frac{\kappa}{m}(N_\gamma - 1)$, with corresponding eigenvector $\boldsymbol{\gamma}/N_\gamma^{1/2} \in B_0(N_\gamma)$. The other eigenvalues are 1, with multiplicity

$p - N_\gamma$, and $1 - \frac{\kappa}{m}$, with multiplicity $N_\gamma - 1$. Hence $\lambda_1(\Sigma_\gamma) - \lambda_2(\Sigma_\gamma) = \frac{\kappa}{m}(N_\gamma - 1)$. Define

$$\Gamma_0 := \left\{ g \in \{0, 1\}^p : \left| N_g - \frac{p\kappa}{m} \right| \leq \frac{k}{20} \right\},$$

where $N_g := \sum_{j=1}^p g_j$. We note that by Bernstein's inequality (e.g. [Shorack and Wellner, 1986](#), p. 855) that

$$\mathbb{P}(\gamma \in \Gamma_0) \geq 1 - 2e^{-k/800}. \quad (13)$$

If $g \in \Gamma_0$, the conditional distribution of $Y_1/\sqrt{750}$ given $\gamma = g$ belongs to $\mathcal{P}_p(n, k, \theta)$ for $\theta \leq \frac{\kappa}{750m}(N_g - 1)$ and all sufficiently large $n \in \mathcal{N}$. By hypothesis, it follows that for $g \in \Gamma_0$,

$$\mathbb{E} \left\{ L(\hat{v}^{(n)}(\mathbf{Y}/\sqrt{750}), v_1(Q_\gamma)) \mid \gamma = g \right\} \leq K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}} \leq 1000 K_0 n^{-\frac{5(1-\alpha)}{2(5-\alpha)}} \sqrt{\log n}.$$

for all $n \in \mathcal{N}$ sufficiently large. Then by Lemma 13 in Appendix C, for $\hat{S}(\cdot)$ defined in Step 3 of Algorithm 3, for $g \in \Gamma_0$, and $n \in \mathcal{N}$ sufficiently large,

$$\begin{aligned} \mathbb{E} \left\{ |S \setminus \hat{S}(\hat{v}^{(n)}(\mathbf{Y}/\sqrt{750}))| \mid \gamma = g \right\} &\leq 2N_g \mathbb{E} \left\{ L(\hat{v}^{(n)}(\mathbf{Y}/\sqrt{750}), v_1(Q_\gamma))^2 \mid \gamma = g \right\} \\ &\leq 2000 K_0 N_g n^{-\frac{5(1-\alpha)}{2(5-\alpha)}} \sqrt{\log n}. \end{aligned}$$

We deduce by Markov's inequality that for $g \in \Gamma_0$, and $n \in \mathcal{N}$ sufficiently large,

$$\mathbb{P} \left\{ |S \cap \hat{S}(\hat{v}^{(n)}(\mathbf{Y}/\sqrt{750}))| \leq 16N_\gamma/17 \mid \gamma = g \right\} \leq 34000 K_0 n^{-\frac{5(1-\alpha)}{2(5-\alpha)}} \sqrt{\log n}. \quad (14)$$

Let

$$\begin{aligned} \Omega_{0,n} &:= \{\gamma \in \Gamma_0\} \cap \{|S \cap \hat{S}(\hat{v}^{(n)}(\mathbf{Y}/\sqrt{750}))| > 16N_\gamma/17\} \\ \Omega'_{0,n} &:= \{\gamma' \in \Gamma_0\} \cap \{|S \cap \hat{S}(\hat{v}^{(n)}(\mathbf{X}/\sqrt{750}))| > 16N_{\gamma'}/17\} =: \Omega'_{1,n} \cap \Omega'_{2,n}, \end{aligned}$$

say, where $N_{\gamma'} := \sum_{j=1}^p \gamma'_j$. When $n \in \mathcal{N}$ is sufficiently large, we have on the event $\Omega'_{0,n}$ that

$$|\{j \in \hat{S}(\hat{v}^{(n)}(\mathbf{X}/\sqrt{750})) : w_j \in K\}| > 3k/4. \quad (15)$$

Now set

$$\Omega'_{3,n} := \left\{ \text{nb}(u, \{w_j : j \in S'\}) \leq \frac{k}{2} \text{ for all } u \in V(G) \setminus K \right\}.$$

Recall the definition of \hat{K} from Step 4 of Algorithm 3. We claim that for sufficiently large $n \in \mathcal{N}$,

$$\Omega'_{0,n} \cap \Omega'_{3,n} \subseteq \{\hat{K} = K\}.$$

To see this, note that for $n \in \mathcal{N}$ sufficiently large, on $\Omega'_{0,n}$ we have $K \subseteq \hat{K}$ by (15). For the reverse inclusion, note that if $u \in V(G) \setminus K$, then on $\Omega'_{0,n} \cap \Omega'_{3,n}$, we have for sufficiently large $n \in \mathcal{N}$ that

$$\begin{aligned} \text{nb}\left(u, \{w_j : j \in \hat{S}(\hat{v}^{(n)}(\mathbf{X}/\sqrt{750}))\}\right) &\leq |\{w_j : j \in \hat{S}\} \setminus K| + \text{nb}(u, \{w_j : j \in \hat{S}\} \cap K) \\ &\leq |\{w_j : j \in \hat{S}\} \setminus K| + \text{nb}(u, \{w_j : j \in S'\}) < \frac{k}{4} + \frac{k}{2} = \frac{3k}{4}. \end{aligned}$$

This establishes our claim. We conclude that for sufficiently large $n \in \mathcal{N}$,

$$\mathbb{P}(\hat{K} \neq K) \leq \mathbb{P}((\Omega'_{0,n} \cap \Omega'_{3,n})^c) \leq \mathbb{P}((\Omega'_{0,n})^c) + \mathbb{P}(\Omega'_{1,n} \cap (\Omega'_{3,n})^c). \quad (16)$$

Now, by Lemma 14, we have

$$|\mathbb{P}(\Omega'_{0,n}) - \mathbb{P}(\Omega_{0,n})| \leq d_{\text{TV}}(\mathcal{L}(\mathbf{X}, \boldsymbol{\gamma}'), \mathcal{L}(\mathbf{Y}, \boldsymbol{\gamma})) \leq \frac{18}{5L}. \quad (17)$$

Moreover, by a union bound and Hoeffding's inequality, for sufficiently large $n \in \mathcal{N}$,

$$\mathbb{P}(\Omega'_{1,n} \cap (\Omega'_{3,n})^c) \leq \sum_{g \in \Gamma_0} \mathbb{P}((\Omega'_{3,n})^c | \boldsymbol{\gamma} = g) \mathbb{P}(\boldsymbol{\gamma} = g) \leq me^{-k/800}. \quad (18)$$

We conclude by (16), (17), (13), (14) and (18) that for sufficiently large $n \in \mathcal{N}$,

$$\mathbb{P}(\hat{K} \neq K) \leq \frac{18}{5L} + 2e^{-k/800} + 34000K_0 n^{-\frac{5(1-\alpha)}{2(5-\alpha)}} \sqrt{\log n} + me^{-k/800} \rightarrow 0$$

as $n \rightarrow \infty$. This contradicts Assumption (A1), and therefore completes the proof. \square

Lemma 7. *Let $g = (g_1, \dots, g_p)^\top \in \{0, 1\}^p$, and let Y_1, \dots, Y_n be independent random vectors, each distributed as $\text{GV}_p^g(\pi_0)$ for some $\pi_0 \in (0, 1/2]$, where the Graph Vector distribution $\text{GV}_p^g(\pi_0)$ is defined in the proof of Theorem 6. For any $u \in B_0(\ell)$, let $V(u) := \mathbb{E}\{(u^\top Y_1)^2\}$ and $\hat{V}(u) := n^{-1} \sum_{i=1}^n (u^\top Y_i)^2$. Then for every $1 \leq \ell \leq 2/\pi_0$, every $n \in \mathbb{N}$ and every $\delta > 0$,*

$$\mathbb{P}\left[\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 750 \max\left\{\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n}\right\}\right] \leq \delta.$$

In other words, $\text{GV}_p^g(\pi_0) \in \text{RCC}_p(\ell, 750)$ for all $\pi_0 \in (0, 1/2]$ and $\ell \leq 2/\pi_0$.

Proof of Lemma 7. We can write

$$Y_i = \xi_i \{(1 - \epsilon_i)R_i + \epsilon_i(g + \tilde{R}_i)\},$$

where ξ_i , ϵ_i and R_i are independent, where ξ_i is a Rademacher random variable, where $\epsilon_i \sim \text{Bern}(\pi_0)$, where $R_i = (r_{i1}, \dots, r_{ip})^\top$ has independent Rademacher coordinates, and where $\tilde{R}_i = (\tilde{r}_{i1}, \dots, \tilde{r}_{ip})^\top$ with $\tilde{r}_{ij} := (1 - g_j)r_{ij}$. Thus, for any $u \in B_0(\ell)$, we have

$$(u^\top Y_i)^2 = (1 - \epsilon_i)(u^\top R_i)^2 + \epsilon_i(u^\top g)^2 + \epsilon_i(u^\top \tilde{R}_i)^2 + 2\epsilon_i(u^\top \tilde{R}_i)(u^\top g).$$

Hence, writing $S := \{j : g_j = 1\}$,

$$\begin{aligned} |\hat{V}(u) - V(u)| &\leq \left| \frac{1}{n} \sum_{i=1}^n (1 - \epsilon_i)(u^\top R_i)^2 - (1 - \pi_0) \right| + \frac{(u^\top g)^2}{n} \left| \sum_{i=1}^n (\epsilon_i - \pi_0) \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i)^2 - \pi_0 \|u_{S^c}\|_2^2 \right| + \left| \frac{2u^\top g}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i) \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n (1 - \epsilon_i) \{ (u^\top R_i)^2 - 1 \} \right| + \frac{1 + (u^\top g)^2 + \|u_{S^c}\|_2^2}{n} \left| \sum_{i=1}^n (\epsilon_i - \pi_0) \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \{ (u^\top \tilde{R}_i)^2 - \|u_{S^c}\|_2^2 \} \right| + \left| \frac{2u^\top g}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i) \right|. \quad (19) \end{aligned}$$

We now control the four terms on the right-hand side of (19) separately. For the first term, note that the distribution of R_i is subgaussian with parameter 1. Writing $N_\epsilon := \sum_{i=1}^n \epsilon_i$, it follows by the same argument as in the proof of Proposition 1(i) that for any $s > 0$,

$$\begin{aligned} &\mathbb{P} \left(\sup_{u \in B_0(\ell)} \left| \frac{1}{n} \sum_{i=1}^n (1 - \epsilon_i) \{ (u^\top R_i)^2 - 1 \} \right| \geq 2s \right) \\ &= \mathbb{E} \left\{ \mathbb{P} \left(\sup_{u \in B_0(\ell)} \left| \frac{1}{n - N_\epsilon} \sum_{i: \epsilon_i=0} \{ (u^\top R_i)^2 - 1 \} \right| \geq \frac{2ns}{n - N_\epsilon} \mid N_\epsilon \right) \right\} \\ &\leq e^9 p^\ell \mathbb{E} \left[\exp \left\{ - \frac{n \left(\frac{ns}{n - N_\epsilon} \right)^2}{4 \left(\frac{ns}{n - N_\epsilon} \right) + 32} \right\} \right] \\ &\leq e^9 p^\ell \exp \left(- \frac{ns^2}{4s + 32} \right). \end{aligned}$$

We deduce that for any $\delta > 0$,

$$\mathbb{P} \left(\sup_{u \in B_0(\ell)} \left| \frac{1}{n} \sum_{i=1}^n (1 - \epsilon_i) \{ (u^\top R_i)^2 - 1 \} \right| \geq 16 \max \left\{ \sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \right\} \right) \leq e^9 \delta. \quad (20)$$

For the second term on the right-hand side of (19), note first that for any $u \in B_0(\ell)$, we have by Cauchy-Schwarz that

$$(u^\top g)^2 \leq \|u_S\|_0 \|u_S\|_2^2 \leq \|u_S\|_0 \leq \ell.$$

We deduce using Bernstein's inequality for Binomial random variables (e.g. [Shorack and Wellner, 1986](#), p. 855) that for any $s > 0$,

$$\begin{aligned} \mathbb{P}\left\{\sup_{u \in B_0(\ell)} \frac{1 + (u^\top g)^2 + \|u_{S^c}\|_2^2}{n} \left| \sum_{i=1}^n (\epsilon_i - \pi_0) \right| \geq s\right\} &\leq \mathbb{P}\left\{\frac{1}{n} \left| \sum_{i=1}^n (\epsilon_i - \pi_0) \right| \geq \frac{s}{3\ell}\right\} \\ &\leq 2 \exp\left(-\frac{ns^2}{18\ell^2\pi_0 + 2s\ell}\right) \leq 2 \max\left\{\exp\left(-\frac{ns^2}{(19 + \sqrt{37})\ell^2\pi_0}\right), \exp\left(-\frac{ns}{(1 + \sqrt{37})\ell}\right)\right\}. \end{aligned}$$

By assumption, $\ell\pi_0 \leq 2$. Hence, for any $\delta > 0$,

$$\begin{aligned} \mathbb{P}\left\{\sup_{u \in B_0(\ell)} \frac{1 + (u^\top g)^2 + \|u_{S^c}\|_2^2}{n} \left| \sum_{i=1}^n (\epsilon_i - \pi_0) \right| \geq (1 + \sqrt{37}) \max\left(\sqrt{\frac{\ell \log(1/\delta)}{n}}, \frac{\ell \log(1/\delta)}{n}\right)\right\} \\ \leq 2\delta. \end{aligned} \quad (21)$$

The third term on the right-hand side of (19) can be handled in a very similar way to the first. We find that for every $\delta > 0$,

$$\mathbb{P}\left(\sup_{u \in B_0(\ell)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \{(u^\top \tilde{R}_i)^2 - \|u_{S^c}\|_2^2\} \right| \geq 16 \max\left\{\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n}\right\}\right) \leq e^9 \delta. \quad (22)$$

For the fourth and final term, from the definition of \tilde{R}_i , we have for any $u \in B_0(\ell)$ that

$$\left| \frac{2u^\top g}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i) \right| \leq \frac{2\ell^{1/2}}{n} \left| \sum_{j:g_j=0} u_j \sum_{i:\epsilon_i=1} r_{ij} \right| \leq \frac{2\ell}{n} \max_{j:g_j=0} \left| \sum_{i:\epsilon_i=1} r_{ij} \right|.$$

Hence by Hoeffding's inequality, for any $s > 0$,

$$\begin{aligned} \mathbb{P}\left\{\sup_{u \in B_0(\ell)} \left| \frac{2(u^\top g)}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i) \right| \geq s\right\} &\leq \mathbb{E}\left\{\mathbb{P}\left(\max_{1 \leq j \leq p} \left| \sum_{i:\epsilon_i=1} r_{ij} \right| \geq \frac{ns}{2\ell} \mid N_\epsilon\right)\right\} \\ &\leq 2p \mathbb{E}\left\{\exp\left(-\frac{n^2 s^2}{8\ell^2 N_\epsilon}\right)\right\} \leq 2p \inf_{t>0} \left\{\exp\left(-\frac{n^2 s^2}{8\ell^2 t}\right) + \mathbb{P}(N_\epsilon > t)\right\} \\ &\leq 2p \inf_{t>0} \left\{\exp\left(-\frac{n^2 s^2}{8\ell^2 t}\right) + \exp\left(-t \log \frac{t}{n\pi_0} + t - n\pi_0\right)\right\}, \end{aligned}$$

where the final line follows by Bennett's inequality (e.g. [Shorack and Wellner, 1986](#), p. 440).

Choosing $t = \max\left(e^2 n\pi_0, \frac{ns}{2^{3/2}\ell}\right)$, we find

$$\begin{aligned} \mathbb{P}\left\{\sup_{u \in B_0(\ell)} \left| \frac{2(u^\top g)}{n} \sum_{i=1}^n \epsilon_i (u^\top \tilde{R}_i) \right| \geq s\right\} \\ \leq 2p \max\left\{\exp\left(-\frac{ns^2}{8e^2\ell^2\pi_0}\right) + \exp\left(-\frac{ns}{2^{3/2}\ell}\right), 2 \exp\left(-\frac{ns}{2^{3/2}\ell}\right)\right\} \\ \leq 4p \max\left\{\exp\left(-\frac{ns^2}{16e^2\ell}\right), \exp\left(-\frac{ns}{2^{3/2}\ell}\right)\right\}. \end{aligned}$$

We deduce that for any $\delta > 0$,

$$\mathbb{P} \left[\sup_{u \in B_0(\ell)} \left| \frac{2(u^\top g)}{n} \sum_{i=1}^n \epsilon_i(u^\top \tilde{R}_i) \right| \geq 4e \max \left\{ \sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \right\} \right] \leq 4\delta. \quad (23)$$

We conclude from (19), (20), (21), (22) and (23) that for any $\delta > 0$,

$$\mathbb{P} \left[\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 750 \max \left\{ \sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \right\} \right] \leq \delta,$$

as required. \square

C Appendix: Ancillary results

We collect here various results used in the proofs in Appendices A and B.

Proposition 8. *Let $P \in \text{RCC}_p(n, \ell, C)$ and suppose that $\ell \log p \leq n$. Then*

$$\mathbb{E}_P \sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \leq \left(1 + \frac{1}{\log p}\right) C \sqrt{\frac{\ell \log p}{n}}.$$

Proof. By setting $\delta = p^{1-t}$ in the RCC condition, we find that

$$\mathbb{P} \left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq C \max \left\{ \sqrt{\frac{t\ell \log p}{n}}, \frac{t\ell \log p}{n} \right\} \right) \leq \min(1, p^{1-t})$$

for all $t \geq 0$. It follows that

$$\begin{aligned} \mathbb{E}_P \sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| &= \int_0^\infty \mathbb{P} \left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq s \right) ds \\ &\leq C \sqrt{\frac{\ell \log p}{n}} + C \sqrt{\frac{\ell \log p}{n}} \int_1^{\frac{n}{\ell \log p}} \frac{1}{2} p^{1-t} t^{-1/2} dt + C \frac{\ell \log p}{n} \int_{\frac{n}{\ell \log p}}^\infty p^{1-t} dt \\ &\leq C \sqrt{\frac{\ell \log p}{n}} \left\{ 1 + \int_1^\infty p^{1-t} dt \right\} = \left(1 + \frac{1}{\log p}\right) C \sqrt{\frac{\ell \log p}{n}}, \end{aligned}$$

as required. \square

Lemma 9. *Let $\epsilon \in (0, 1/2)$, let $\ell \in \{1, \dots, p\}$ and let $A \in \mathbb{R}^{p \times p}$ be symmetric. Then there exists $\mathcal{N}_\epsilon \subseteq B_0(\ell)$ with $|\mathcal{N}_\epsilon| \leq \binom{p}{\ell} \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ such that*

$$\sup_{u \in B_0(\ell)} |u^\top A u| \leq (1 - 2\epsilon)^{-1} \max_{u \in \mathcal{N}_\epsilon} |u^\top A u|.$$

Proof. Let $\mathcal{I}_\ell := \{I \subseteq \{1, \dots, p\} : |I| = \ell\}$, and for $I \in \mathcal{I}_\ell$, let $B_I := \{u \in B_0(\ell) : u_{I^c} = 0\}$. Thus

$$B_0(\ell) = \bigcup_{I \in \mathcal{I}_\ell} B_I.$$

For each $I \in \mathcal{I}_\ell$, by Lemma 10 of [Kim and Samworth \(2014\)](#), there exists $\mathcal{N}_{I,\epsilon} \subseteq B_0(\ell)$ such that $|\mathcal{N}_{I,\epsilon}| \leq \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ and such that for any $x \in B_I$, there exists $x' \in \mathcal{N}_{I,\epsilon}$ with $\|x - x'\| \leq \epsilon$. Let $u_I \in \operatorname{argmax}_{u \in B_I} |u^\top A u|$ and find $v_I \in \mathcal{N}_{I,\epsilon}$ such that $\|u_I - v_I\| \leq \epsilon$. Then

$$|u_I^\top A u_I| \leq |v_I^\top A v_I| + |(u_I - v_I)^\top A v_I| + |u_I^\top A (u_I - v_I)| \leq \max_{u \in \mathcal{N}_{I,\epsilon}} |u^\top A u| + 2\epsilon |u_I^\top A u_I|.$$

Writing $\mathcal{N}_\epsilon := \bigcup_{I \in \mathcal{I}_\ell} \mathcal{N}_{I,\epsilon}$, we note that $|\mathcal{N}_\epsilon| \leq \binom{p}{\ell} \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ and that

$$\sup_{u \in B_0(\ell)} |u^\top A u| = \max_{I \in \mathcal{I}_\ell} \sup_{u \in B_I} |u^\top A u| \leq (1 - 2\epsilon)^{-1} \max_{I \in \mathcal{I}_\ell} \max_{u \in \mathcal{N}_{I,\epsilon}} |u^\top A u| = (1 - 2\epsilon)^{-1} \max_{u \in \mathcal{N}_\epsilon} |u^\top A u|,$$

as required. \square

Lemma 10 (Variant of the Gilbert–Varshamov Lemma). *Let $\alpha, \beta \in (0, 1)$ and $k, p \in \mathbb{N}$ be such that $k \leq \alpha\beta p$. Writing $\mathcal{S} := \{x = (x_1, \dots, x_p)^\top \in \{0, 1\}^p : \sum_{j=1}^p \mathbb{1}_{\{x_j=1\}} = k\}$, there exists a subset \mathcal{S}_0 of \mathcal{S} such that for all distinct $x = (x_1, \dots, x_p)^\top, y = (y_1, \dots, y_p)^\top \in \mathcal{S}_0$, we have $\sum_{j=1}^p \mathbb{1}_{\{x_j \neq y_j\}} \geq 2(1 - \alpha)k$ and such that*

$$\log |\mathcal{S}_0| \geq \rho k \log(p/k),$$

where $\rho := \frac{\alpha}{-\log(\alpha\beta)}(-\log \beta + \beta - 1)$.

Proof. See [Massart \(2007, Lemma 4.10\)](#). \square

Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$. Recall that if P is absolutely continuous with respect to Q , we define the Kullback–Leibler divergence between P and Q to be $D(P\|Q) := \int_{\mathcal{X}} \log(dP/dQ) dP$, where dP/dQ denotes the Radon–Nikodym derivative of P with respect to Q . If P is not absolutely continuous with respect to Q , we set $D(P\|Q) := \infty$.

Lemma 11 (Generalised Fano’s Lemma). *Let P_1, \dots, P_M be probability distributions on a measurable space $(\mathcal{X}, \mathcal{B})$, and assume that $D(P_i\|P_j) \leq \beta$ for all $i \neq j$. Then any measurable function $\hat{\psi} : \mathcal{X} \rightarrow \{1, \dots, M\}$ satisfies*

$$\max_{1 \leq i \leq M} P_i(\hat{\psi} \neq i) \geq 1 - \frac{\beta + \log 2}{\log M}.$$

Proof. See [Yu \(1997, Lemma 3\)](#). \square

Lemma 12. Suppose $P \in \mathcal{P}$ and $X_1, \dots, X_n \stackrel{iid}{\sim} P$. Let $\Sigma := \int_{\mathbb{R}^p} xx^\top dP(x)$ and $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$. If $V(u) = \mathbb{E}\{(u^\top X_1)^2\}$ and $\hat{V}(u) := n^{-1} \sum_{i=1}^n (u^\top X_i)^2$ for $u \in B_0(2)$, then

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq 2 \sup_{u \in B_0(2)} |\hat{V}(u) - V(u)|.$$

Proof. Let e_r denote the r th standard basis vector in \mathbb{R}^p and write $X_i = (X_{i,1}, \dots, X_{i,p})^\top$. Then

$$\begin{aligned} \|\hat{\Sigma} - \Sigma\|_\infty &= \max_{r,s \in \{1, \dots, p\}} \left| \frac{1}{n} \sum_{i=1}^n (X_{i,r} X_{i,s}) - \mathbb{E}(X_{1,r} X_{1,s}) \right| \\ &\leq \max_{r,s \in \{1, \dots, p\}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \left(\frac{1}{2} e_r + \frac{1}{2} e_s \right)^\top X_i \right\}^2 - \mathbb{E} \left[\left\{ \left(\frac{1}{2} e_r + \frac{1}{2} e_s \right)^\top X_1 \right\}^2 \right] \right| \\ &\quad + \max_{r,s \in \{1, \dots, p\}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \left(\frac{1}{2} e_r - \frac{1}{2} e_s \right)^\top X_i \right\}^2 - \mathbb{E} \left[\left\{ \left(\frac{1}{2} e_r - \frac{1}{2} e_s \right)^\top X_1 \right\}^2 \right] \right| \\ &\leq 2 \sup_{u \in B_0(2)} |\hat{V}(u) - V(u)|, \end{aligned}$$

as required. \square

Lemma 13. Let $v = (v_1, \dots, v_p)^\top \in B_0(k)$ and let $\hat{v} = (\hat{v}_1, \dots, \hat{v}_p)^\top \in \mathbb{R}^p$ be such that $\|\hat{v}\|_2 = 1$. Let $S := \{j \in \{1, \dots, p\} : v_j \neq 0\}$. Then for any $\hat{S} \in \operatorname{argmax}_{1 \leq j_1 < \dots < j_k \leq p} \sum_{r=1}^k |\hat{v}_{j_r}|$, we have

$$L(\hat{v}, v)^2 \geq \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2.$$

Proof. By the Cauchy–Schwarz inequality, and then by definition of \hat{S} ,

$$\begin{aligned} 1 - L(\hat{v}, v)^2 &= \left(\sum_{j \in S \setminus \hat{S}} \hat{v}_j v_j + \sum_{j \in S \cap \hat{S}} \hat{v}_j v_j \right)^2 \leq \left(2 \sum_{j \in S \setminus \hat{S}} \hat{v}_j^2 + \sum_{j \in S \cap \hat{S}} \hat{v}_j^2 \right) \left(\frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2 + \sum_{j \in S \cap \hat{S}} v_j^2 \right) \\ &\leq \left(\sum_{j \in S \setminus \hat{S}} \hat{v}_j^2 + \sum_{j \in S \setminus \hat{S}} \hat{v}_j^2 + \sum_{j \in S \cap \hat{S}} \hat{v}_j^2 \right) \left(1 - \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2 \right) \leq 1 - \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2, \end{aligned}$$

as required. \square

Recall the definition of the total variation distance d_{TV} given in the proof of [Theorem 6](#).

Lemma 14. Let X and Y be random elements taking values in a measurable space (F, \mathcal{F}) , and let (G, \mathcal{G}) be another measurable space.

(a) If $\phi : F \rightarrow G$ is measurable, then

$$d_{\text{TV}}(\mathcal{L}(\phi(X)), \mathcal{L}(\phi(Y))) \leq d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)).$$

(b) Let Z be a random element taking values in (G, \mathcal{G}) , and suppose that Z is independent of (X, Y) . Then

$$d_{\text{TV}}(\mathcal{L}(X, Z), \mathcal{L}(Y, Z)) = d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)).$$

Proof. (a) For any $A \in \mathcal{G}$, we have

$$|\mathbb{P}\{\phi(X) \in A\} - \mathbb{P}\{\phi(Y) \in A\}| = |\mathbb{P}\{X \in \phi^{-1}(A)\} - \mathbb{P}\{Y \in \phi^{-1}(A)\}| \leq d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)).$$

Since $A \in \mathcal{G}$ was arbitrary, the result follows.

(b) Define $\phi : F \times G \rightarrow F$ by $\phi(w, z) := w$. Then ϕ is measurable, and using the result of part (a),

$$d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = d_{\text{TV}}(\mathcal{L}(\phi(X, Z)), \mathcal{L}(\phi(Y, Z))) \leq d_{\text{TV}}(\mathcal{L}(X, Z), \mathcal{L}(Y, Z)).$$

For the other inequality, let \mathcal{A} denote the set of subsets A of $\mathcal{F} \otimes \mathcal{G}$ with the property that given $\epsilon > 0$, there exist sets $B_{1,F}, \dots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \dots, B_{n,G} \in \mathcal{G}$ such that, writing $B := \cup_{i=1}^n (B_{i,F} \times B_{i,G})$, we have $\mathbb{P}((X, Z) \in A \Delta B) < \epsilon$ and $\mathbb{P}((Y, Z) \in A \Delta B) < \epsilon$. Here, the binary operator Δ denotes the symmetric difference of two sets, so that $A \Delta B := (A \cap B^c) \cup (A^c \cap B)$. Note that $\mathcal{F} \times \mathcal{G} \subseteq \mathcal{A}$. Now suppose $A \in \mathcal{A}$ so that, given $\epsilon > 0$, we can find sets $B_{1,F}, \dots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \dots, B_{n,G} \in \mathcal{G}$ with the properties above. Observe that we can write

$$B^c = \bigcup_{I \subseteq \{1, \dots, n\}} \left(\bigcap_{i \in I} B_{i,F}^c \times \bigcap_{i \in I} B_{i,G} \cap \bigcap_{i \in I^c} B_{i,G}^c \right).$$

For each $I \subseteq \{1, \dots, n\}$, the sets $\cap_{i \in I} B_{i,F}^c$ belong to \mathcal{F} , and $\{\cap_{i \in I} B_{i,G} \cap \cap_{i \in I^c} B_{i,G}^c : I \subseteq \{1, \dots, n\}\}$ is a family of disjoint sets in \mathcal{G} . Moreover,

$$\mathbb{P}((X, Z) \in A^c \Delta B^c) = \mathbb{P}((X, Z) \in A \Delta B) < \epsilon,$$

and similarly $\mathbb{P}((Y, Z) \in A^c \Delta B^c) < \epsilon$. We deduce that $A^c \in \mathcal{A}$. Finally, if (A_n) is a disjoint sequence in \mathcal{A} , then let $A := \cup_{n=1}^\infty A_n$, and given $\epsilon > 0$, find $m \in \mathbb{N}$ such that $\mathbb{P}((X, Z) \in A \setminus \cup_{i=1}^m A_i) < \epsilon/2$ and $\mathbb{P}((Y, Z) \in A \setminus \cup_{i=1}^m A_i) < \epsilon/2$. Now, for each $i = 1, \dots, m$,

find sets $B_{i1,F}, \dots, B_{in_i,F} \in \mathcal{F}$ and disjoint sets $B_{i1,G}, \dots, B_{in_i,G} \in \mathcal{G}$ such that, writing $B_i := \cup_{j=1}^{n_i} (B_{ij,F} \times B_{ij,G})$, we have $\mathbb{P}((X, Z) \in A_i \triangle B_i) < \epsilon/(2m)$ and $\mathbb{P}((Y, Z) \in A_i \triangle B_i) < \epsilon/(2m)$. It is convenient to relabel the sets $\{(B_{ij,F}, B_{ij,G}) : i = 1, \dots, m, j = 1, \dots, n_i\}$ as $\{(C_{1,F}, C_{1,G}), \dots, (C_{N,F}, C_{N,G})\}$, where $N := \sum_{i=1}^m n_i$. This means that we can write

$$\bigcup_{i=1}^m B_i = \bigcup_{k=1}^N (C_{k,F} \times C_{k,G}) = \bigcup_{K \subseteq \{1, \dots, N\}, K \neq \emptyset} \left(\bigcup_{k \in K} C_{k,F} \times \bigcap_{k \in K} C_{k,G} \cap \bigcap_{k \in K^c} C_{k,G}^c \right).$$

Now, for each non-empty subset K of $\{1, \dots, N\}$, the set $\cup_{k \in K} C_{k,F}$ belongs to \mathcal{F} , and $\{\cap_{k \in K} C_{k,G} \cap \cap_{k \in K^c} C_{k,G}^c : K \subseteq \{1, \dots, N\}, K \neq \emptyset\}$ is a family of disjoint sets in \mathcal{G} . Moreover,

$$\mathbb{P}((X, Z) \in A \triangle \bigcup_{i=1}^m B_i) \leq \sum_{i=1}^m \mathbb{P}((X, Z) \in A_i \triangle B_i) + \frac{\epsilon}{2} < \epsilon,$$

and similarly, $\mathbb{P}((Y, Z) \in A \triangle \bigcup_{i=1}^m B_i) < \epsilon$. We deduce that $A \in \mathcal{A}$, so \mathcal{A} is a σ -algebra containing $\mathcal{F} \times \mathcal{G}$, so \mathcal{A} contains $\mathcal{F} \otimes \mathcal{G}$.

Now suppose that $A \in \mathcal{F} \otimes \mathcal{G}$. By the argument above, given $\epsilon > 0$, there exist sets $B_{1,F}, \dots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \dots, B_{n,G} \in \mathcal{G}$ such that $\mathbb{P}((X, Z) \in A \triangle \bigcup_{i=1}^n (B_{i,F} \times B_{i,G})) < \epsilon/2$ and $\mathbb{P}((Y, Z) \in A \triangle \bigcup_{i=1}^n (B_{i,F} \times B_{i,G})) < \epsilon/2$. It follows that

$$\begin{aligned} |\mathbb{P}((X, Z) \in A) - \mathbb{P}((Y, Z) \in A)| &\leq \sum_{i=1}^n |\mathbb{P}(X \in B_{i,F}, Z \in B_{i,G}) - \mathbb{P}(Y \in B_{i,F}, Z \in B_{i,G})| + \epsilon \\ &= \sum_{i=1}^n \mathbb{P}(Z \in B_{i,G}) |\mathbb{P}(X \in B_{i,F}) - \mathbb{P}(Y \in B_{i,F})| + \epsilon \\ &\leq d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) + \epsilon. \end{aligned}$$

Since $A \in \mathcal{A}$ and $\epsilon > 0$ were arbitrary, we conclude that

$$d_{\text{TV}}(\mathcal{L}(X, Z), \mathcal{L}(Y, Z)) \leq d_{\text{TV}}(\mathcal{L}(X), \mathcal{L}(Y)),$$

as required. □

D Appendix: A brief introduction to computational complexity theory

The following is intended to give a short introduction to notions in computational complexity theory referred to in the paper. A good reference for further information is [Arora and Barak \(2009\)](#), from which much of the following is inspired.

A *computational problem* is the task of generating a desired output based on a given input. Formally, defining $\{0, 1\}^* := \cup_{k=1}^{\infty} \{0, 1\}^k$ to be the set of all finite strings of zeros and ones, we can view a computational problem as a function $F : \{0, 1\}^* \rightarrow \mathcal{P}(\{0, 1\}^*)$, where $\mathcal{P}(A)$ denotes the power set of a set A . The interpretation is that $F(s)$ describes the set of acceptable output strings (solutions) for a particular input string s .

Loosely speaking, an *algorithm* is a collection of instructions for performing a task. Despite the widespread use of algorithms in mathematics throughout history, it was not until 1936 that Alonzo Church and Alan Turing formalised the notion by defining notational systems called the λ -calculus and Turing machines respectively (Church, 1936; Turing, 1936). Here we define an algorithm to be a *Turing machine*:

Definition 1. A Turing machine M is a pair (Q, δ) , where

- Q is a finite set of states, among which are two distinguished states q_{start} and q_{halt} .
- δ is a ‘transition’ function from $Q \times \{0, 1, \sqcup\}$ to $Q \times \{0, 1, \sqcup\} \times \{L, R\}$.

A Turing Machine can be thought of as having a reading head that can access a tape consisting of a countably infinite number of squares, labelled $0, 1, 2, \dots$. When the Turing machine is given an input $s \in \{0, 1\}^*$, the tape is initialised with the components of s in its first $|s|$ tape squares (where $|\cdot|$ denotes the length of a string in $\{0, 1\}^*$) and with ‘blank symbols’ \sqcup in its remaining squares. The Turing machine starts in the state $q_{\text{start}} \in Q$ with its head on the 0th square and operates according to its transition function δ . When the machine is in state $q \in Q$ with its head over the i th tape square that contains the symbol $a \in \{0, 1, \sqcup\}$, and if $\delta(q, a) = (q', a', L)$, the machine overwrites a with a' , updates its state to q' , and moves to square $i - 1$ (or to square $i + 1$ if the third component of the transition function is R instead of L). The Turing machine stops if it reaches state $q_{\text{halt}} \in Q$ and outputs the vector of symbols on the tape before the first blank symbol. If the Turing machine M terminates (in finitely many steps) with input s , we write $M(s)$ for its output.

We say an algorithm (Turing machine) M *solves a computational problem* F if M terminates for every input $s \in \{0, 1\}^*$, and $M(s) \in F(s)$. A computational problem is *solvable* if there exists a Turing machine that solves it. It turns out that other notions of an algorithm (including Church’s λ -calculus and modern computer programming languages) are equivalent in the sense that the set of solvable problems is the same.

A *polynomial time algorithm* is a Turing machine M for which there exist $a, b > 0$ such that for all input strings $s \in \{0, 1\}^*$, M terminates after at most $a|s|^b$ transitions. We say a problem F is *polynomial time solvable*, written $F \in \mathbf{P}$, if there exists a polynomial time algorithm that solves it⁵.

A *nondeterministic Turing machine* has the same definition as that for a Turing machine except that the transition function δ becomes a set-valued function $\delta : Q \times \{0, 1, \sqcup\} \rightarrow \mathcal{P}(Q \times \{0, 1, \sqcup\} \times \{L, R\})$. The idea is that, while in state q with its head over symbol a , a nondeterministic Turing machine replicates $|\delta(q, a)|$ copies of itself (and its tape) in the current configuration, each exploring a different possible future configuration in the set $\delta(q, a)$. Each replicate branches to further replicates in the next step. The process continues until one of its replicates reaches the state q_{halt} . At that point, the Turing machine replicate that has halted outputs its tape content and all replicates stop computation. A *nondeterministic polynomial time algorithm* is a nondeterministic Turing machine M_{nd} for which there exist $a, b > 0$ such that for all input strings $s \in \{0, 1\}^*$, M_{nd} terminates after at most $a|s|^b$ steps. (We count all replicates of M_{nd} making one parallel transition as one step.) We say a computational problem F is *nondeterministically polynomial time solvable*, written $F \in \mathbf{NP}$, if there exists a nondeterministic polynomial time algorithm that solves it⁶.

Clearly $\mathbf{P} \subseteq \mathbf{NP}$, but it is not currently known if these classes are equal. It is widely believed that $\mathbf{P} \neq \mathbf{NP}$, and many computational lower bounds for particular computational problems have been proved conditional under this assumption. Working under this hypothesis, a common strategy is to relate the algorithmic complexity of one computational problem to another. We say a computational problem F is *polynomial time reducible* to another problem G , written as $F \leq_{\mathbf{P}} G$, if there exist polynomial time algorithms M_{in} and M_{out} such that $M_{\text{out}} \circ G \circ M_{\text{in}}(s) \subseteq F(s)$. In other words, $F \leq_{\mathbf{P}} G$ if we can convert an input of F to an input of G through M_{in} , and translate every solution of G back to a solution for F through M_{out} .

Definition 2. A computational problem G is *NP-hard* if $F \leq_{\mathbf{P}} G$ for all $F \in \mathbf{NP}$. It is *NP-complete* if it is in \mathbf{NP} and is *NP-hard*.

⁵In fact, some authors write \mathbf{FP} (short for ‘Functional Polynomial Time’) for the class we have denoted as \mathbf{P} here. The notation \mathbf{P} is then reserved for the subset of computational problems consisting of so-called *decision problems* F , where $F(s) \in \{\{0\}, \{1\}\}$ for all $s \in \{0, 1\}^*$.

⁶Again, some authors write \mathbf{FNP} for the class we have denoted as \mathbf{NP} here.

Karp (1972) showed that a large number of natural computational problems are NP-complete, including the Clique problem mentioned in Section 4. The Turing machines and nondeterministic Turing machines introduced above are both non-random. In some situations (e.g. statistical problems), it is useful to consider random procedures:

Definition 3. A probabilistic Turing machine M_{pr} is a triple (Q, δ, X) , where

- Q is a finite set of states, among which are two distinguished states q_{start} and q_{halt} .
- δ is a transition function from $Q \times \{0, 1, \sqcup\} \times \{0, 1\}$ to $Q \times \{0, 1, \sqcup\} \times \{L, R\}$.
- $X = (X_1, X_2, \dots)$ is an infinite sequence of independent $\text{Bern}(1/2)$ random variables.

In its t th step, if a probabilistic Turing machine M_{pr} is in state q with its reading head over symbol a , and $\delta(q, a, X_t) = (q', a', L)$, then M_{pr} overwrites a with a' , updates its state to q' and moves its reading head to the left (or to the right if $\delta(q, a, X_t) = (q', a', R)$). A *randomised polynomial time algorithm* is a probabilistic Turing machine M_{pr} for which there exist $a, b > 0$ such that for any $s \in \{0, 1\}^*$, M_{pr} terminates in at most $a|s|^b$ steps. We say a computational problem F is *solvable in randomised polynomial time*, written as $F \in \text{BPP}$, if, given $\epsilon > 0$, there exists a randomised polynomial time algorithm $M_{\text{pr}, \epsilon}$ such that $\mathbb{P}(M_{\text{pr}, \epsilon}(s) \in F(s)) \geq 1 - \epsilon$.

In the above discussion, the classes P, NP, BPP are all defined through worst-case performance of an algorithm, since we require the time bound to hold for every input string s . However, in many statistical applications, the input string s is drawn from some distribution \mathcal{D} on $\{0, 1\}^*$, and it is the average performance of the algorithm, rather than the worst case scenario, that is of more interest. We say such a random problem is solvable in randomised polynomial time if, given $\epsilon > 0$, there exists a randomised polynomial time algorithm $M_{\text{pr}, \epsilon}$ such that, when $s \sim \mathcal{D}$, independent of X , we have $\mathbb{P}(M_{\text{pr}}(s) \in F(s)) \geq 1 - \epsilon$. Note that the probability here is taken over both the randomness in s and the randomness in X . Similar to the non-random cases, we can talk about randomised polynomial time reduction. If M_F is a randomised polynomial time algorithm for a computational problem F , then $M_{\text{out}} \circ M_F \circ M_{\text{in}}$ is a potential randomised polynomial time algorithm for another problem G for suitably constructed randomised polynomial time algorithms M_{in} and M_{out} . One such construction is the key to the proof of Theorem 6.

References

- Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R., and Xie, N. (2007) Testing k -wise and almost k -wise independence. *Proceedings of the thirty-ninth ACM Symposium on Theory of Computing*, 496–505.
- Alon, N., Krivelevich, M. and Sudakov, B. (1998) Finding a large hidden clique in a random graph. *Proceedings of the ninth annual ACM-SIAM Symposium on Discrete Algorithms*, 594–598.
- Ames, B. P. W. and Vavasis, S. A. (2011) Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, **129**, 69–89.
- Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.
- Applebaum, B., Barak, B. and Wigderson, A. (2010) Public-key cryptography from different assumptions. *Proceedings of the forty-second ACM Symposium on Theory of Computing*, 171–180.
- Arora, S. and Barak, B. (2009) *Computational Complexity: A Modern Approach*. Cambridge University Press, Cambridge.
- Bach, F., Ahipaşaoğlu, S. D., d’Aspremont, A. (2010) Convex relaxations for subset selection. Available at <http://arxiv.org/abs/1006.3601>.
- Baik, J., Ben Arous, G. and Péché, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.
- Berthet, Q. (2014) Optimal testing for planted satisfiability problems. Available at <http://arxiv.org/abs/1401.2205>.
- Berthet, Q. and Rigollet P. (2013) Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41**, 1780–1815.
- Berthet, Q. and Rigollet P. (2013) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. W&CP*, **30**, 1046–1066.

- Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, **41**, 1055–1084.
- Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, **41**, 3074–3110.
- Chandrasekaran, V. and Jordan, M. I. (2013) Computational and statistical tradeoffs via convex relaxation. *Proc. Nat. Acad. Sci.*, **110**, E1181–E1190.
- Chen, Y. and Xu, J. (2014) Chen, Y. and Xu, J. (2014) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. Available at <http://arxiv.org/abs/1402.1267>.
- Church, A. (1936) An unsolvable problem of elementary number theory. *Amer. J. Math.*, **58**, 345–363.
- d’Aspremont, A. El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.
- Davis, C. & Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7**, 1–46.
- Diaconis, P. and Freedman D. (1980) Finite exchangeable sequences. *Ann. Probab.*, **8**, 745–764.
- Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. S. and Xiao, Y. (2013) Statistical algorithms and a lower bound for detecting planted cliques. *Proceedings of the forty-fifth annual ACM Symposium on Theory of Computing*, 655–664.
- Feige, U. and Krauthgamer, R. (2000) Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms*, **16**, 195–208.
- Feige, U. and Krauthgamer, R. (2003) The probable value of the Lovász–Schrijver relaxations for a maximum independent set. *SIAM J. Comput.*, **32**, 345–370.
- Feige, U. and Ron, D. (2010) Finding hidden cliques in linear time. *Discrete Math. Theor. Comput. Sci. Proc.*, 189–204.

- Feldman, V., Perkins, W. and Vempala, S. (2013) On the Complexity of Random Satisfiability Problems with Planted Solutions. Available at <http://arxiv.org/abs/1311.4821>.
- Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland.
- Grimmett, G. R. and McDiarmid C. J. H. (1975) On colouring random graphs. *Math. Proc. Cambridge Philos. Soc.*, **77**, 313–324.
- Hajek, B., Wu, Y. and Xu, J. (2014) Computational lower bounds for community detection on random graphs. Available at <http://arxiv.org/abs/1406.6625>.
- Hazan, E. and Krauthgamer, R. (2011) How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, **40**, 79–91.
- Jerrum, M. (1992) Large cliques elude the Metropolis process. *Random Structures Algorithms*, **3**, 347–359.
- Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.
- Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component technique based on the LASSO, *J. Comput. Graph. Statist.*, **12**, 531–547.
- Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010) Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, **11**, 517–553.
- Juels, A. and Peinado, M. (2000) Hiding cliques for cryptographic security. *Des. Codes Cryptogr.*, **20**, 269–280.
- Karp, R. M. (1972) Reducibility among combinatorial problems. In R. E. Miller et al. (Eds.), *Complexity of Computer Computations*, 85–103. Springer, New York.
- Kim, A. K.-H. and Samworth R. J. (2014) Global rates of convergence in log-concave density estimation. Available at <http://arxiv.org/abs/1404.2298>.
- Kučera, L. (1995) Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, **57**, 193–212.

- Lanczos, C. (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, **45**, 255–282.
- Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.
- Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, **41**, 772–801.
- Ma, Z. and Wu, Y. (2013) Computational barriers in minimax submatrix detection. Available at <http://arxiv.org/abs/1309.5914>.
- Massart, P. (2007) *Concentration Inequalities and Model Selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII - 2003*. Springer, Berlin/Heidelberg.
- Nemirovski, A. (2004) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.* **15**, 229–251.
- Nesterov, Yu. (2005) Smooth minimization of nonsmooth functions. *Math. Program., Ser. A*, **103**, 127–152.
- Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, **17**, 1617–1642.
- Shen, D., Shen, H. and Marron, J. S. (2013) Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.*, **115**, 317–333.
- Shorack, G. R. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics*. Wiley, New York.
- Turing, A. (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, **2**, 230–265.
- van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

- van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.
- Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013) Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA. *Advances in Neural Information Processing Systems (NIPS)*, **26**.
- Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.
- Wang, Z., Lu, H. and Liu, H. (2014) Nonconvex statistical optimization: minimax-optimal Sparse PCA in polynomial time. *In preparation*.
- Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.
- Yu, B. (1997) Assouad, Fano and Le Cam. In Pollard, D., Torgersen, E. and Yang G. L. (Eds.) *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, 423–435. Springer, New York.
- Yu, Y., Wang, T. and Samworth, R. J. (2014) A useful variant of the Davis–Kahan theorem for statisticians. Available at <http://arxiv.org/abs/1405.0680>.
- Yuan, X.-T. and Zhang, T. (2013) Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, **14**, 899–925.
- Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *J. Mach. Learn. Res. W&CP*, **35**, 921–948.
- Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal components analysis. *J. Comput. Graph. Statist.* **15**, 265–86.